The Transnational Technology Law Review

Artificial Intelligence and the Law: Special Essay Collection

Critical Approaches to Data, Algorithms and Science in the Law

Editors
Awotula Damilola
Effoduh Jake Okechukwu
Yu Annie

https://transnationaltechlawreview.com

Vol. 1, Oct. 2025

The Transnational Technology Law Review

Artificial Intelligence and the Law: Special Essay Collection

Critical Approaches to Data, Algorithms and Science in the Law

https://transnationaltechlawreview.com

Editors
Awotula Damilola
Effoduh Jake Okechukwu
Yu Annie

Recommended Collection Citation: Awotula, Effoduh, and Yu (eds*), Critical Approaches to Data, Algorithms and Science in the Law: Special Essay Collection* (The Transnational Technology Law Review, Vol. 1, Oct. 2025).

Sample Chapter Citation: Boyd C*, Exploitation Through Algorithmic Training: The Struggle To Protect Creativity In The AI Era* in Awotula, Effoduh, and Yu (eds), *Critical Approaches to Data, Algorithms and Science in the Law: Special Essay Collection* (The Transnational Technology Law Review, Vol. 1, Oct. 2025).

Preface

Artificial intelligence is reshaping fundamental aspects of society (employment, commerce, speech regulation, justice, and warfare), while simultaneously amplifying racial inequalities, eroding privacy, and delegating moral judgment to algorithms. These challenges strike at the law's core mandate to safeguard fairness, accountability, and human dignity.

This collection presents some student reflections on these emerging issues, mostly drawn from short papers submitted for JUR 204: Critical Approaches to Data, Algorithm, and Science in the Law at Lincoln Alexander School of Law, Toronto Metropolitan University, during the 2024/2025 academic year under Professor Jake Effoduh's tuition.

We envision this as the inaugural volume in an annual tradition, creating a platform for students and budding researchers to offer their perspectives on the technologies shaping their futures. The essays engage diverse contemporary challenges: algorithmic bias in recruitment and sentencing, technology's perpetuation of racial inequalities, content moderation politics, intellectual property exploitation, privacy vulnerabilities, and even autonomous weapons. Across these domains emerges a unifying thesis: AI cannot be understood apart from the social structures it reflects and transforms.

These contributions are modest attempts to compel not merely through their topical relevance, but through their distinctive budding perspectives. As digital natives entering the legal profession, these students approach technology and law with their own critical imagination, situating doctrine within broader struggles over power, marginalization, and freedom. Their analyses demonstrate that governing emerging technologies requires both technical literacy and a deep understanding of justice and equity, and this is part of their own attempt at the same.

Contributors

Bakshi Simran is a JD candidate at the Lincoln Alexander School of Law and graduated with an Honours Specialization in Biology from Western University in 2022. She continues to pursue research opportunities which prioritize the mobilization of scientific knowledge into policymaking.

Basten Nicole is a JD candidate at the Lincoln Alexander School of Law at Toronto Metropolitan University. She hopes to pursue a legal career in securities regulations, and her current research focuses on consumer protection and information privacy.

Boyd Charlotte is a JD candidate at Lincoln Alexander School of Law, Toronto Metropolitan University. Charlotte holds a Bachelor of Arts Honours in Political Studies, Economics, and Philosophy from Queen's University. She has a passion for innovation and creative industries and loves to play pickleball in her free time.

Butt Amna is a JD candidate at the Lincoln Alexander School of Law. She completed her undergraduate studies at Brock University, specializing in economics and finance. Through her work, Amna seeks to explore and advance critical analyses of the finance industry deploying various legal theory lenses.

Capano Marisa is a JD candidate with hands-on experience in business law, legal research, and advocacy. With a background in both client-facing legal work and systemic reform initiatives, she's interested in how law interacts with data, design, and public policy to shape more inclusive legal systems.

Elkhinovich Lawrence is a JD candidate at the Lincoln Alexander School of Law at Toronto Metropolitan University. His interests include litigation, technology governance, cross-border transactions, and student welfare. He has experience in legal research, advocacy, contract drafting, and policy analysis. Lawrence has also worked with community organizations to promote equitable access to legal resources.

Gill Rajvir is a JD candidate at Lincoln Alexander School of Law, passionate about criminal law and social justice. With experience in legal research, client advocacy, and community engagement, she has contributed to research projects aimed at dismantling systemic injustices. Rajvir is dedicated to fostering equitable legal outcomes.

Malik Shizza is a JD candidate at the Lincoln Alexander School of Law, with a background in Psychology and Criminology from the University of Toronto. Her interests include exploring the evolving field of health law, particularly its intersections with modern technology and artificial intelligence, including patient rights, healthcare ethics and regulatory frameworks.

Marr Julia is a JD Candidate with lived experiences in the welfare system. Julia's community inspired her to attend law school and pursue a career in family law. In addition to law school, Julia is a legal clinic casework at the Lincoln Alexander School of Law legal clinic and Halton Community Legal Services.

McColl Grace holds a BCom from the Smith School of Business and is a JD Candidate at the Lincoln Alexander School of Law. Upon graduation, Grace will be clerking at the Ontario Court of Appeal and then returning to a Toronto law firm as Corporate Associate. Her academic interests include public international law and feminist legal theory.

Pabla Sunny was born and raised in Windsor, Ontario, and combines business leadership with a passion for advocacy. With a Bachelor of Commerce and an MBA, he is also a licensed realtor, investor, and restaurant owner. Now a JD candidate focused on litigation, Sunny draws on his entrepreneurial background and extensive community involvement to inform a results-driven approach to the law.

Patel Angeli is a second year JD candidate at the Lincoln Alexander School of Law. She holds a B.A. (Honours) from Carleton University and is passionate about exploring the intersection of law and AI, particularly in the business and privacy law sectors.

Peterson L Erin is a JD candidate at the Lincoln Alexander School of Law, specializing in the intersection of AI and law. With a background in Computer Engineering and AI, she explores the legal and ethical implications of artificial intelligence, focusing on effective regulatory frameworks for emerging technologies.

Srirajasingam Charanija a JD candidate at the Lincoln Alexander School of Law, and holds a Bachelor's in Legal Studies and Business, with a minor in Human Resource Management from the University of Waterloo. As the daughter of Tamil immigrants, she is passionate about integrating her heritage into her legal work.

Zhang Helen is a Canadian-born Chinese student interested in the intersection of criminal law, evidence accuracy, and technology. She is a JD candidate at the Lincoln Alexander School of Law. In her downtime, Helen likes to relax with her cat and play soccer.

Table of Contents

Chapter 1

Smart Farming, Not Smart for Smallholders: The Marginalization of Small-Scale Farmers in the Age of Big Data and AgriTech

Simran Bakshi (she/her)

JD Candidate, Lincoln Alexander School of Law

Abstract

The dominant techno-optimistic narrative surrounding Agriculture 4.0 posits that agritechnologies ("AgriTech") will democratize agricultural production, enhance food security, and improve the livelihoods of smallholder farmers. However, this vision often overlooks the concentration of corporate power in agrotechnology and the reinforcement of existing inequalities within the global agri-food system. This short paper critically examines the extent to which AgriTech marginalizes smallholder farmers through biased data practices, the alienation of farm labour, and the erosion of local knowledge networks. Using a Marxist approach to a Science and Technology Studies ("STS") and Critical Data Studies framework, this paper argues that AgriTech serves as a strategic tool for agribusinesses to entrench their dominance in global agriculture while exacerbating the vulnerabilities of smallholder farmers. The analysis focuses on (1) data and algorithmic biases that prioritize large-scale industrial farming over smallholder and alternative farming systems, (2) the alienation of farmers from decision-making processes through digital locks and algorithmic governance, and (3) the erosion of collective knowledge exchange through the commodification of agricultural information. Given that smallholder farmers operate 84% of the world's farms, addressing these asymmetries is crucial for achieving an equitable agricultural future. This paper concludes by proposing policy interventions, including the creation of an Agricultural Data Privacy and Transparency Act and the inclusion of smallholder farmer perspectives in AgriTech governance. By critically assessing the socio-political dimensions of AgriTech, this paper highlights the urgent need for policies that challenge corporate monopolization and promote sustainable, anti-capitalistic agricultural development.

Keywords: AgriTech, Smallholder Farmers, Algorithmic Bias, Marxist, Agricultural Marginalization

Introduction

'The windmill gives you society with the feudal lord; the steam mill, society with the industrial capitalist' - Karl Marx, The Poverty of Philosophy, 1847[1]

A dominant techno-optimistic narrative known as 'agriculture 4.0'[2] provides that agritechnologies ("AgriTech") will be a revolutionary force for the global agro-food system's pressing challenges. Scholars suggest that agritechnologies wield the power to "democratize" the agricultural sector by providing further accessibility to smallholder farmers. For instance, the Food and Agriculture Organization of the United Nations ("FAO") claimed that effective data management would improve the livelihood of smallholder farmers[3] by mitigating informational asymmetries and financial barriers.[4] However, this optimistic vision often overlooks corporate and political power concentration within agrotechnology. This is evidenced by historical so-called revolutionary agricultural waves, like the Green Revolution, which has been shown to widen the gap between small and large-scale farmers.[5] The increasing concentration of corporate power in the global agri-food system has paved the way for the same companies dominating seeds, agrochemicals, fertilizers, and farm equipment, like John Deere and Bayer-Monsanto, to expand market control into the agrotechnology sector through investments in data-intensive farm management platforms. This raises critical questions about these agritechnologies' beneficiaries, losers, and underlying polito-social interests.

The increasing turn to artificial intelligence ("AI") driven AgriTech systems that are managed by a few, but powerful AgriTech companies, raises even more cause for worries. Given that by a 2019 FAO report, 84% of the world's 608 million farms are smallholdings,[6] it is crucial to examine whether smallholder farmers' narratives and knowledge systems are marginalized in developing Big Data Analytics and AI within agrotechnology.

Hence, I explore what forms of knowledge are associated with Big Data Analytics and AI within agrotechnology and whether the narratives of smallholder farmers are marginalized by their development. Drawing on Marxist, Science and Technology Studies and Critical Data Studies, I argue that AI driven AgriTech companies act to reinforce their dominant position in global agri-food systems[7] while jeopardizing smallholder farmers and the environment.

Specifically, I argue that AI -driven agritechnologies work to further reproduce asymmetric power relations: (1) through data and algorithmic biases, (2) the alienation of farmer labour, and (3) the erosion of smallholder farmers' local knowledge networks. As some scholars have rightly argued, these elements allow agrotechnology corporations to reinforce productivist logic and create platform-related dependencies.

---

[1] Paul D'Amato, "Marxism and the Making of History" (17 April 2014), online: SocialistWorker <socialistworker.org/2014/04/17/the-making-of-history> [perma.cc/4LHN-7R5W].
[2] David Christian Rose & Jason Chilvers, "Agriculture 4.0: Broadening Responsible Innovation in an Era of Smart Farming" (2018) 2:87 Frontiers in Sustainable Food Systems 1 – 7.
[3] Smallholder farmers are defined as farmers dependent on their own labour and resources to produce food and predominately operated farms 2 hectares in size or less.
[4] Food and Agriculture Organization of the United Nations, The State of Food and Agriculture 2021. Making agrifood systems more resilient to shocks and stresses (last visited 24 March 2025), online (PDF): < openknowledge.fao.org/server/api/core/bitstreams/1e61f82a-618c-467a-a37f-545580094a1d/content>.
[5] Prabhu L Pingali, "Green Revolution: Impacts, limits, and the path ahead" (2012) 109:31 Proceedings National Academy Sciences 12302.
[6] Sarah K Lowder, Jakob Skoet & Terri Raney, "The Number, Size, and Distribution of Farms, Smallholder Farms, and Family Farms Worldwide" (2016) 87 World Development 27.
[7] Global agrifood systems is defined as, "contemporary knowledge, institutions, infrastructures, practices and crops that define the predominant patterns of food production and consumption."

Data and Algorithmic Biases: Robbery of the Worker and Soil

'All progress in capitalist agriculture is a progress in the art, not only of robbing the worker, but of robbing the soil.' - Karl Marx[8]

a. Data and Algorithmic Biases

Data and algorithmic biases in agrotechnologies perpetuate asymmetrical power relations between smallholder and large-scale farmers and threaten agricultural biodiversity. First, smart farming technologies' datasets and AI models often contain biases that disproportionately benefit large-scale commercial farms at the expense of smallholder and alternative farmers. For instance, a 2025 quantitative research indicates that current technological developments are biased toward crop-specific pest and pathogen threats in high-income, industrialized countries.[9] Moreover, data-driven recommendations provided by large agri-business-owned technological platforms, like Climate FieldView, uphold capitalist agricultural practices, which scholars frame as "efforts to preserve [agricultural oligopolies] by monetizing [farmers'] behaviour."[10] For example, agrotechnologies have predominantly focused on creating recommendations for farming practices (i.e., irrigation system practices, seed selection, fertilizer application) solely for commodity crops (i.e., corn, soybeans) cultivated on medium to large farms.[11]

Another example is N-Manager, an analytics tool developed under the FarmCommand platform.[12] Designed to optimize nitrogen fertilizer use, the tool is explicitly tailored for farms using chemical inputs, excluding organic and low-input systems that smallholder and alternative farmers use. This bias is further compounded by N-Manager's reliance on a crop modelling system, Decision Support System for Agrotechnology Transfer ("DSSAT"), primarily focused on commodity crops like corn, soybeans, and rice. Although DSSAT models over 170 cultivars of corn, it offers limited support for crops common in smallholder farming, like diversified crops and fruit trees.[13] Such AI technologies designed and optimized for large-scale, commercial production implicitly favour large agribusinesses that possess adequate capital and are incongruent with the needs of smallholder and organic farmers that grow diversified crops and have limited capital.

b. Increased Vulnerability of Agricultural Systems

In addition to the exploitation of labour, Marx also considered the extent to which the raw materials of production are degraded.[14] As large-scale commercial farms implement agricultural technology to boost productivity under capitalistic business models, they also increase monocultures, resulting in low biodiversity systems, particularly in the Global South.[15]

---

[8] John Bellamy Foster & Brett Clark, "The Robbery of Nature: Capitalism and the Metabolic Rift" (01 July 2018), online: <https://monthlyreview.org/2018/07/01/the-robbery-of-nature/> [perma.cc/S9K8-E63F].
[9] Jacob Moscona & Karthik A Sastry, "Inappropriate Technology: Evidence from Global Agriculture" (2022) Harvard University, Working Paper No 3886019, online: <papers.ssrn.com/sol3/papers.cfm?abstract_id=3886019>.
[10] Christopher Miles, "The combine will tell the truth: On precision agriculture and algorithmic rationality" (2019) 6:1 Big Data & Society 8.
[11] Maaz Gardezi & Ryan Stock, "Growing algorithmic governmentality: Interrogating the social construction of trust in precision agriculture" (2021) 84 J Rural Studies 5.
[12] Kelly Bronson, Sarah Rotz & Adrian D'Alessandro, "The Human Impact of Data Bias and the Digital Agricultural Revolution" in Harvey S. James, Jr. ed *Handbook on The Human Impact of Agriculture* (Edward Elgar Publishing, 2021) 126.
[13] *Ibid*.
[14] Mailyn Fidler, "Preferring Rabbits To Revolution: A Comparative Analysis of Marxist and Local Food Movement Critiques of Capitalist Agriculture" (2012) 5 Intersect, online (pdf): <ojs.stanford.edu/ojs/index.php/intersect/article/view/313/179>.
[15] Andres Suarez & Wencke Gwozdz, "On the relation between monocultures and ecosystem services in the Global South: A review" (2023) 278 Biological Conservation 109870.

Agribusinesses' cultivation of such commodified crops relies heavily on environmentally harmful synthetic fertilizers, increasing crop vulnerability to diseases and climate change.[16]

While some sub-Saharan African scholars argue that despite these biases, AgriTech can still enhance the ecological productivity of smallholder farms, and their deployment should not be restricted solely to large, commercial farmers, they also note that technological mismatches between smallholder farming systems and 'productivist' agricultural practices can significantly reduce agricultural productivity, especially when AgriTech is poorly suited to local conditions.[17] The study demonstrated that smallholder farmers in less productive regions are less likely to adopt improved AgriTech, such as genetically modified seeds, due to the 'inappropriateness' of the technology offered for the farmers' needs. It was found that the AgriTech was most inappropriate for local conditions, and smallholder farmers were significantly less likely to use improved seeds.[18] The study also suggests that seed adoption could be as much as 14% higher than current levels in the absence of this mismatch.[19]

However, for farmers in low-productivity areas, access to and the suitability of improved inputs remains a significant barrier. Hence, large agribusinesses continue marginalizing smallholder farmers by strategically adapting AgriTech to suit local conditions, reinforcing a technological divide favouring large-scale industrial farming over more diverse, smallholder-driven agricultural systems.

'Pulling-Away of the Natural Ground' and the Alienation of Labour

Despite the potential of AgriTech to address the labour shortage issue,[20] it exacerbates the alienation of agricultural labour, diminishing smallholder farmer autonomy and perpetuating AgriTech corporations' 'productivist' logic. In Marxist terms, alienation refers to the alienation of farm labour from the ownership of the means of production.[21] For instance, Carolan posits that the trust farmers instill in data-driven recommendations and their acceptance of 'disciplinary directives offered by algorithmic authority'[22] leads to their subservience to a form of governance by the algorithm. AgriTech corporations have built moral trust between smallholder farmers and smart farming technologies by positioning their knowledge products as superior to their local experiential knowledge of their labour.[23] Consequently, smallholder farmers have internalized the authority of AI algorithms in shaping their decision-making.[24]

Furthermore, AgriTech corporations use license agreements to gain control over farmers' data and form digital "locks." For instance, certain agreements prevent farmers from gaining access to information that would assist in repairing their AgriTech.[25] From a Marxist lens, these digital "locks" are equivalent to alienation, where farmers are increasingly alienated from the means of production—the tools and knowledge necessary for their work. A recent example is an ongoing

---

[16] Gerardo Patron-Cano, *Modern capitalism and food commoditization: The limitations of industrial agriculture and the challenges of sustainable alternatives*. (Masters Thesis, University of Denver, 2015) [unpublished].
[17] Abdulai Adams & Emmanuel Tetteh Jumpah, "Agricultural technologies adoption and smallholder farmers' welfare: Evidence from Northern Ghana" (2021) 9:1 Cogent Econs & Finance 2.
[18] *Supra* note 17 at 3.
[19] *Ibid*.
[20] A Subeesh & CR Mehta, "Automation and digitization of agriculture using artificial intelligence and internet of things" (2021) 5 Artificial Intelligence in Agriculture 278.
[21] Supra note 1.
[22] Michael Carolan, "Automated agrifood futures: Robotics, labor and the distributive politics of digital agriculture" (2020) 47:1 J Peasant Studies 184.
[23] *Supra* note 11 at 4.
[24] *Ibid*.
[25] Helena Shilomboleni et al, "Scaling up innovations in smallholder agriculture: Lessons from the Canadian international food security research fund" (2019) 175 Agricultural Systems 58.

lawsuit against John Deere where a farmer complained about restricted access to operating software of their farm equipment, making it impossible for them to self-diagnose fault or use third party repairers.[26] By controlling access to vital resources, AgriTech corporations intensify the exploitation of smallholder farmers, shifting power away from the producers and consolidating wealth and market control for oligopolies.

Erosion of Local Knowledge Networks

AgriTech undermines the local, experiential knowledge that traditionally empowers smallholder farmers. This diminishing of autonomy is critical as it is considered a social tool to identify, mitigate, navigate, and translate the experiences of being a farmer in the broader network of agricultural relations.[27] First, the opaque algorithms of AgriTech impede smallholder farmers' abilities to experiment and innovate, a core element of the environmental learning process.[28] Furthermore, the influx of automated AgriTech, like automatic irrigation systems and harvesting robots, rooted in capitalistic logic, impedes the collective exchange of wisdom and collaboration between farmers on a local scale.

This aligns with Marx's concept of the "pulling away of the natural ground" in agriculture, where machinery, fertilizers, and seeds detached farmers from traditional knowledge and practices. While some digital technologies may complement farmers' knowledge, agrotechnologies seek to 'pull away' from the local, experiential learning processes traditionally grounded in smallholder farming practices. As these technologies dictate specific farming operations through opaque algorithms, they inhibit farmers' ability to experiment and innovate, which is critical for environmental learning and adaptation.[29] Furthermore, the individualistic nature of AgriTech can create division amongst farmers, making it challenging for them to learn from one another and work together as they have traditionally done. Instead of sharing knowledge and testing new ideas together, farmers who use AgriTech may become more isolated, relying on AgriTech platforms instead of engaging in local, community-based learning and support.[30]

Recommendations: Collectivism and Inclusion of Unheard Voices

Given the exacerbation of asymmetrical power relations by agrotechnologies, assessing how policies must be restructured to protect smallholder farmers from the dominance of agri-digital giants is critical. To combat a capitalistic agricultural model, Marx calls for revolutionary actions "against the existing social and political order of things."[31] This ideology for transformation invites us to abolish the prevailing capitalist agricultural model, consider policies and legal frameworks that empower smallholder farmers, and challenge corporate control over the global agri-food system. The following solutions focus on collective farming approaches and integrating sustainable, inclusive technologies to address these systemic imbalances.

a. Knowledge in Technology Development and Everyday Encounters

The absence of smallholder farmer perspectives in the design and governance of AgriTech further exacerbates the power asymmetry between smallholder and large-commercial farmers. As such,

[26] Alina Selyukh, John Deere faces U.S. lawsuit over farmers' ability to repair tractors (last visited 15 March 2025) online: <https://www.npr.org/2025/01/15/nx-s1-5260895/john-deere-ftc-lawsuit-right-to-repair-tractors>.
[27] Paul V Stock & Jérémie Forney, "Farmer autonomy and the farming self" (2014) 36 J Rural Studies 160.
[28] Glenn Davis Stone, "Towards a General Theory of Agricultural Knowledge Production: Environmental, Social, and Didactic Learning" (2016) 38:1 J of Culture & Agriculture 5.
[29] *Ibid*.
[30] *Supra* note 28 at 17.
[31] Jodi Dean, "The Communist Manifesto: An idea whose time has come again" (01 July 2018), online: <www.plutobooks.com/blog/communist-manifesto-idea-whose-time-come-again/> [perma.cc/S9K8-E63F].

when designing policy instruments, it is critical to shift the narrative beyond techno-optimist discourse that often emphasizes AgriTech's 'revolutionary' impact and prioritizes large-scale commodity farmers at the expense of small-scale and alternative farmers.[32] This narrative limits policymakers from focusing on market control for AgriTech corporations rather than acknowledging issues like farmer and environmental justice.[33] As such, governance regimes of sustainable AgriTech must involve the substantive inclusion of diverse stakeholder communities.[34]

The proposition of "everyday encounters"[35] is a potential solution to resist the dominant capitalistic and techno-optimistic narrative of rapid AgriTech implementations. "Everyday encounters" consider how farmers engage with technology daily and assess how smallholder farmers' worldviews influence their attitude towards technology use.[36] This is particularly important as incorporating the voices of these 'harder to reach'[37] community members, like Indigenous, organic, and alternative farmers, often serves as a challenge for policymakers operating in traditional processes.[38] This may also inform how data can be disseminated for the collective benefit of stakeholders beyond large-scale commercial farmers. By implementing "everyday encounters," AgriTech developers can consider the expertise of smallholder and alternative farmers and the uncertain nature of crops in their data collection and algorithm design methods.[39] As such, AgriTech developers will likely be encouraged to consider the needs of smallholder or alternative farmers when designing AgriTech instead of strategically prioritizing agrotechnology corporations.

b. Agricultural Data Privacy and Transparency Act

The power asymmetries and risks posed by the digital divide in AgriTech necessitate legal frameworks that address data privacy and transparency while prioritizing smallholder farmers. Protecting agricultural data could be a viable legal intervention. Farmers should be granted the right to ease of data portability when they choose to switch AgriTech, and should be advised on how, what and the usage of their collected agricultural data. Data ownership is also critical. For instance, the 2024 Privacy and Security Principles for Farm Data by the American Farm Bureau Federation noted data ownership by farmers as a fundamental principle.

In the context of Canada, parliament could pass an Agricultural Data Privacy and Transparency Act, which would explicitly classify agricultural data—particularly geospatial and agronomic data—as personal information, ensuring its protection under existing data privacy laws like the Personal Information Protection and Electronic Documents Act ("PIPEDA") in Canada.[40] Canadian privacy laws like PIPEDA and the Privacy Act focus primarily on personal information identifying individuals.[41]

---

[32] David Christian Rose et al, "The old, the new, or the old made new? Everyday counter-narratives of the so-called fourth agricultural revolution" (2023) 40:2 Agriculture and Human Values 423.
[33] Emily Duncan et al, "New but for whom? Discourses of innovation in precision agriculture" (2021) 38 Agriculture and Human Values 1181.
[34] Kelly Bronson, "Looking through a responsible innovation lens at uneven engagements with digital farming" (2019) 90 NJAS-Wageningen J Life Sciences 100294.
[35] *Supra* note 32 at 424.
[36] *Ibid*.
[37] Paul Hurley et al, "Co-designing the environmental land management scheme in England: The why, who and how of engaging 'harder to reach' stakeholders" (2022) 4:3 People & Nature 744.
[38] Auvikki de Boon, Camilla Sanstrom & David Christian Rose, "Governing agricultural innovation: A comprehensive framework to underpin sustainable transitions" (2022) 89 J Rural Studies 407.
[39] Laura Foster et al, "Smart farming and artificial intelligence in East Africa: Addressing indigeneity, plants, and gender" (2023) 3 Smart Agricultural Technology 100132.
[40] *Personal Information Protection and Electronic Documents Act*, SC 2000, c 5.
[41] *Privacy Act*, RSC 1985, c P-21.

However, they fail to account for the specific nature of agricultural data, which can include geospatial data that indirectly identifies farmers. This critical gap creates ambiguity around how agricultural data should be protected and managed, leaving farmers vulnerable to exploitation by AgriTech corporations. By treating agricultural data as personal information, the proposed Act would ensure informed consent from farmers, requiring precise, explicit agreements on how their data is collected, used, and shared by agricultural technology providers. Other countries should do the same to protect farmers.

Conclusion

Despite the potential of AgriTech to advance 'productivist' agriculture, large agribusinesses' techno-optimistic narrative surrounding AgriTech will lead to more significant power asymmetries between large and smallholder farmers. Especially given the emergence of non-agricultural actors, like IBM,[42] in the AgriTech industry, it is crucial to implement policies and regulations to protect the exploitation of smallholder and alternative farmers. The future of digital farming must prioritize inclusivity and sustainability, not just capital at the expense of smallholder farmers.

[42] Swathi Kumari H & KT Verramanju, "Revolutionizing Agriculture: A Case Study of IBM's AI Innovations" (2023) 7:4 Intl J Applied Engineering & Management 96.

References

Legislation

Personal Information Protection and Electronic Documents Act, Criminal Code, SC 2000, c 5.

Privacy Act, RSC 1985, c P-21.

Articles

Adams, Abdulai & Emmanuel Tetteh Jumpah, "Agricultural technologies adoption and smallholder farmers' welfare: Evidence from Northern Ghana" (2021) 9:1 Cogent Econs & Finance 1 – 19.

Auvikki de Boon, Camilla Sanstrom & David Christian Rose, "Governing agricultural innovation: A comprehensive framework to underpin sustainable transitions" (2022) 89 J Rural Studies 407.

Carolan, Michael "Automated agrifood futures: Robotics, labour and the distributive politics of digital agriculture" (2020) 47:1 J Peasant Studies 184.

Duncan, Emily et al, "New but for whom? Discourses of innovation in precision agriculture" (2021) 38 Agriculture and Human Values 1181.

Fidler, Mailyn, "Preferring Rabbits To Revolution: A Comparative Analysis of Marxist and Local Food Movement Critiques of Capitalist Agriculture" (2012) 5 Intersect 1 – 16.

Foster, Laura et al, "Smart farming and artificial intelligence in East Africa: Addressing indigeneity, plants, and gender" (2023) 3 Smart Agricultural Technology 100132.

Gardezi, Maaz & Ryan Stock, "Growing algorithmic governmentality: Interrogating the social construction of trust in precision agriculture" (2021) 84 J Rural Studies 1 – 11.

Hurley, Paul et al, "Co-designing the environmental land management scheme in England: The why, who and how of engaging 'harder to reach' stakeholders" (2022) 4:3 People & Nature 744.

Lowder, Sarah K Jakob Skoet & Terri Raney, "The Number, Size, and Distribution of Farms, Smallholder Farms, and Family Farms Worldwide" (2016) 87 World Development 16 – 29.

Miles, Christopher, "The combine will tell the truth: On precision agriculture and algorithmic rationality" (2019) 6:1 Big Data & Society 1- 12.

Pingali, Prabhu L, "Green Revolution: Impacts, limits, and the path ahead" (2012) PNAS 109:31 12302 – 12308.

Rose, David Christian et al, "The old, the new, or the old made new? Everyday counter-narratives of the so-called fourth agricultural revolution" (2023) 40:2 Agriculture and Human Values 423.

Rose, David Christian & Jason Chilvers, "Agriculture 4.0: Broadening Responsible Innovation in an Era of Smart Farming" (2018) 2:87 Frontiers in Sustainable Food Systems 1 – 7.

Shilomboleni, Helena et al, "Scaling up innovations in smallholder agriculture: Lessons from the Canadian international food security research fund" (2019) 175 Agricultural Systems 58.

Stock, Paul V & Jérémie Forney, "Farmer autonomy and the farming self" (2014) 36 J Rural Studies 160.

Stone, Glenn Davis, "Towards a General Theory of Agricultural Knowledge Production: Environmental, Social, and Didactic Learning" (2016) 38:1 J of Culture & Agriculture 5.

Suarez, Andres & Wencke Gwozdz, "On the relation between monocultures and ecosystem services in the Global South: A review" (2023) 278 Biological Conservation 109870.

Subeesh, A & CR Mehta, "Automation and digitization of agriculture using artificial intelligence and internet of things." (2021) 5 Artificial Intelligence in Agriculture 278.

Swathi Kumari H & KT Verramanju, "Revolutionizing Agriculture: A Case Study of IBM's AI Innovations" (2023) 7:4 Intl J Applied Engineering & Management 96.

Secondary Material: Working Papers

Moscona, Jacob & Karthik A Sastry, "Inappropriate Technology: Evidence from Global Agriculture" (2022) Harvard University, Working Paper No 3886019, online: <papers.ssrn.com/sol3/papers.cfm?abstract_id=3886019>.

Patron-Cano, Gerardo, *Modern capitalism and food commoditization: The limitations of industrial agriculture and the challenges of sustainable alternatives*. (Masters Thesis, University of Denver, 2015) [unpublished], online (pdf): <digitalcommons.du.edu/cgi/viewcontent.cgi?article=1496&context=etd>.

Secondary Materials: Books

Bronson, Kelly, Sarah Rotz & Adrian D'Alessandro, *Handbook on the human impact of agriculture* (Edward Elgar Publishing, 2021), ch 8 at 126, online (pdf): <sarahrotz.com/wp-content/uploads/2021/06/08-chapter-8_rotz-and-bronson_2021.pdf>.

Other Materials

D'Amato, Paul, "Marxism and the Making of History" (17 April 2014), online: <socialistworker.org/2014/04/17/the-making-of-history> [perma.cc/4LHN-7R5W].

Dean, Jodi, "The Communist Manifesto: An idea whose time has come again" (01 July 2018), online: <www.plutobooks.com/blog/communist-manifesto-idea-whose-time-come-again/> [perma.cc/S9K8-E63F].

Food and Agriculture Organization of the United Nations, The State of Food and Agriculture 2021. Making agrifood systems more resilient to shocks and stresses (FAO, 2021), online (pdf): <https://openknowledge.fao.org/server/api/core/bitstreams/125b023c-002f-4387-9150-dc7fbbd86cbc/content>.

Foster, John Bellamy & Brett Clark, "The Robbery of Nature: Capitalism and the Metabolic Rift" (01 July 2018), online: <https://monthlyreview.org/2018/07/01/the-robbery-of-nature/> [perma.cc/S9K8-E63F].

Chapter 2

Reconciling Privacy and AI in the Digital Age: A Critical Analysis of AI Governance in Canada

Nicole Basten

JD Candidate, Lincoln Alexander School of Law

Abstract

This paper explores the privacy issues in the governance and regulation of Artificial Intelligence ("AI") in Canada. Privacy is an important, multifaceted, and interconnected issue in discussions of data, algorithms, and AI technologies. Privacy has long been recognized as a fundamental right in Canada, yet attempts at its regulation have resulted in a patchwork of *ad hoc* solutions that leave critical gaps in individual protection. Further complicating the matter is the intersection of privacy concerns with the rise of AI, making it difficult to regulate privacy both separately and jointly. Reconciliation of privacy and AI in the digital era must focus on finding the appropriate balance between individual privacy protection without compromising technological advancement. This paper posits that Bill C-27, and its proposed *Artificial Intelligence and Data Act* ("*AIDA"*) is an insufficient solution to AI regulation in Canada. Drawing from a law and economics perspective to critique the objective of the Bill, the conclusions of this work are twofold. First, Bill C-27 overemphasizes innovation and development and leaves privacy in the wake. Second, the ambiguity purposefully built into the *AIDA* may work to reinforce harmful power dynamics, which may result in dangerous consequences for marginalized populations. Further, this paper notes the facade of privacy protection evident in the Bill and underscores the need for strictly proactive privacy protection measures, rather than loosely reactive measures.

Keywords: Privacy, Artificial Intelligence, law and economics, Regulation

Introduction

A recent focus in Canadian legal scholarship is the development of a framework that properly situates artificial intelligence ("AI") within the schema of Canadian law and society. One major point of conflict in discussions about the use of data, algorithms, and AI in Canadian law are the issues that arise specifically with respect to informational privacy. Bill C-27, *The Digital Charter Implementation Act*, was introduced to the House of Commons in 2022 and currently stands as Canada's attempt to create a comprehensive and agile regulatory framework to govern digital privacy, data protection, and AI.

While the future of the Bill[1] is unsettled, it nonetheless reflects the only federal regulatory intervention with regards to AI in Canada and ultimately aims to modernize privacy regulation and promote responsible AI creation and innovation. However, from a law and economics lens,[2] it becomes clear that Bill C-27 prioritizes technological innovation and economic interests and peripheralizes privacy protection. I argue that, if adopted, Bill C-27's proposed *Artificial Intelligence and Data Act* ("*AIDA"*) will act as a hindrance to individual privacy interests in Canada, reinforcing discriminatory power imbalances and supporting profit over privacy.

Bill C-27 Overemphasizes Technological Innovation and Development

The thrust of a law and economic analysis asks whether law delivers economic efficiency. In this context, we must ask whether Bill C-27 properly reconciles privacy with other competing values. In this regard, Bill C-27, then, must balance the commercial interests pursued in technological advancement[3]. The cycle is vicious: technology companies are steadily devising innovative ways to extract more data than ever[4], while "weak, industry-friendly laws"[5] are on the rise as countries strive to maintain technological dominance. Since data is the life wire of AI systems, some argue that privacy interests are in serious jeopardy. This sort of weak regulatory approach is embedded into Bill C-27, whereby it trades off strong privacy protection for technological efficiency and commercial interests. Three aspects built into Bill C-27 are illustrative of how the Bill *prima facie* emphasizes innovation and advancement over privacy interests:

a. Preamble and Purpose

The preamble of Bill C-27 explicitly states that trust in the digital and data-driven economy is key to ensuring its growth. While scholars like Neil Richards posit that "a sustainable and ethical digital society will depend on the trust that is safeguarded by the rules protecting our privacy"[6], if they are to effectively promote a trust relationship, these information rules must weigh privacy interests equally and informatively against economic interests. The *Consumer Privacy Protection Act* ("*CPPA*"), one of the statutes within Bill C-27, states its purpose as the promotion of electronic

---

[1] Bill C-27, An Act to enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act, 1st Sess, 44th Parl, 2022 (at consideration in committee in the House of Commons). [*Bill C-27*]

[2] Lewis Kornhauser "The Economic Analysis of Law", *The Stanford Encyclopedia of Philosophy* (Spring 2022 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2022/entries/legal-econanalysis/>.

[3] Charles I. Jones, and Christopher Tonetti. 2020. "Nonrivalry and the Economics of Data." *American Economic Review*, 110 (9): 2819–58.

[4] Neil Richards, *Why Privacy Matters* (New York: Oxford Press 2022) at 17-34 at p.207; Melissa Heikkilä, "AI companies are finally being forced to cough up for training data" (July 2, 2024) online: MIT Technology Review <https://www.technologyreview.com/2024/07/02/1094508/ai-companies-are-finally-being-forced-to-cough-up-for-training-data/>.

[5] Suzanne Smalley, State privacy laws have been crippled by big tech, new report says (February 1, 2024) online: The Record <https://therecord.media/state-privacy-laws-big-tech-lobbying-report>.

[6] Richards, *supra* note 4 at pp. 8-9.

commerce by protecting personal information that is collected, used, or disclosed in the course of commercial activities. While there is nothing wrong with steering a legislation to meet specific objectives, it also should properly consider other crucial interests.

Although section 5 of the CPPA recognizes "the right to privacy of individuals" it equally recognizes the need of organizations to collect, use or disclose personal information…", but it is silent on which interest takes priority. In short, CPPA tacitly recognizes that these are competing interests but nevertheless fails to reconcile them by not stating expressly that privacy is a fundamental interest. Notwithstanding that Canadian courts have consistently recognized privacy as paramount, giving it a "quasi-constitutional status"[7], it is important to signal expressly to technology companies that privacy interests trump their thirst for data extraction.

Moreover, in light of the fact that technological evolution occurs rapidly and unexpectedly, privacy interests are often *activated* by the misuse of personal data and are not often prone to anticipation.[8] Strong enforcement actions should therefore be easily accessible. However, section 107 of the *CPPA*, which provides a private right of action to individuals who experience contraventions to the *Act*, is contingent on proof of *actual* harm or injury and also requires the action to be sanctioned by the Commissioner.

These sort of safety valves implicitly favour technology companies and explicitly suggest that Bill C-27 mainly seeks to promote and protect technological and economical interests. The Bill's focus on technological advancement over privacy interests thus becomes clear: focused on the profit-making interests of private corporations, the Bill makes no proactive effort to prevent possible personal information and state clearly the paramountcy of privacy interest, leaving open the possibility of unforeseeable privacy incursions. This also highlights the facade of privacy protection hidden in the Bill–a reactionary, post-mortem measure of protection that stands in stark contrast to the constitutionally entrenched notion that privacy protections must be proactive.[9]

b. Compliance is Monitored by the Minister of Innovation

Indicative of the Bill's focus on innovation over privacy, Bill C-27 is overseen, implemented, and enforced by the Minister of Innovation, Science, and Industry (the "Minister"). While the Minister does work alongside the Privacy Commissioner and other government entities, the Minister, whose objective is ultimately the advancement of Canada's economic growth and development, is responsible for the administration and enforcement of the *CPPA*[10]. This poses an interesting tension between the Minister's experience, background, and overall goal of development and innovation and his responsibility to protect the privacy of individuals, which stands in direct opposition to those economic advancement interests. The risk that the application and enforcement of privacy protections within the Bill will be overshadowed by the pursuit of technological advancement is high, thus outweighing the benefits. Moreover, the Bill has been criticized for failing to provide a clear definition of the Minister's role, and for its failure to create independent regulatory bodies to oversee data and AI.[11] Considering that the enforcement of Bill C-27 and its provisions will be the determining factor of whether the Bill properly protects privacy, oversight by the Minister of

[7] Vanessa A MacDonnell, "A Theory of Quasi-Constitutional Legislation", 2016 53-2 *Osgoode Hall Law Journal* 508, 2016 CanLIIDocs 4295 at p. 510.
[8] Matt Malone, A Multidisciplinary Perspective for Policy-Makers, Auditors, and Users, 1st ed (Florida: CRC Press) 73 at p. 72.
[9] *Hunter v Southam*, [1984] 2 S.C.R. 145
[10] Bill C-27, supra note 1 at *Consumer Privacy Protection Act* s. 2.
[11]Teresa Scassa, "Regulating AI in Canada: A Critical Look at the Proposed Artificial Intelligence and Data Act" (2023) 101:1 Canadian Bar Review 1.

Innovation, Science and Industry itself raises questions about the legitimacy of the privacy protections afforded under Bill C-27.[12]

c. *AIDA Fails to Properly Acknowledge the Intersection of AI and Privacy*

AI regulation will not be sufficient unless it properly acknowledges, considers, and specifically aims to protect user informational privacy. The proposed *AIDA* has a focus on AI systems, setting standards for ethical use, accountability, and transparency for the purpose of regulating international and interprovincial trade[13]. While s. 4(b) of the *AIDA* claims to "prohibit certain conduct in relation to artificial intelligence systems that may result in serious harm to individuals or harm to their interests", scholars like Teresa Scassa highlight the ambiguity present in the *AIDA*'s provisions.[14]

For instance, Scassa argues that critical terms, such as "high-impact" remain undefined, to be later delineated by the Minister by virtue of *AIDA* ss. 5(1).[15] The impact of these ambiguities on privacy interests is severe and may result in privacy being foregone in pursuit of AI systems' development. Since the governance of algorithms is currently "played out on an *ad hoc* basis across sectors"[16], the *AIDA* should have made an attempt to facilitate uniform and comprehensive privacy provisions directly related to the creation, dissemination, and use of AI technologies.

These three issues can therefore be used as illustrative examples leading to the conclusion that, *prima facie*, Bill C-27 and the composition of statutes within it appear to emphasize innovation and technological advancement over privacy interests. Each of these issues is also illustrative of the power dynamics reinforced by the legislation, covered in the next section.

Ambiguity Built Into the Legislation Reinforces Harmful Power Dynamics

AI requires at least two constituent elements: algorithms and data.[17] "Data" could mean digital records which, in some instances, may include personal information.[18] Databases are filled with information and then used to train AI programs, which in turn predict "everything from traffic patterns to the location of undocumented migrants".[19] Scholars have rightly stated that, "privacy is about power".[20] The use and distribution of personal information–*data*–is a critical piece of our increasingly digital society. Privacy law expert Neil Richards likens this data to oil of the industrial age, "human information is the fuel of the information economy".[21] The privacy as power debate therefore equates the ability to exploit personal information as social power, and the commodification of personal data cements the power dynamics between individuals and the governments and corporations that gather and distribute the information. Richards proclaims that

[12] *Ibid* at p.4.
[13] Bill C-27, *supra* note 1 at *Artificial Intelligence and Data Act* at s. 4.
[14] *Ibid.*, at s. 1.
[15] Scassa, *supra* note 11 at p. 4.
[16] Joan Donovan et al, "Algorithmic Accountability: A Primer" (April 18, 2018) online: Data & society < https://www.datasociety.net/wp-con- tent/uploads/2018/04/Data_Society_Algorithmic_Accountability_Primer_FINAL.pdf> at p.11.
[17] Bariffi, Francisco Jose, "Artificial Intelligence, Human Rights and Disability" (2021) 26:2 Pensar - Revista de Ciências Jurídicas 1 at p.4.
[18] Vincent C. Müller, "Ethics of Artificial Intelligence and Robotics", *The Stanford Encyclopedia of Philosophy* (Fall 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/fall2023/entries/ethics-ai/>
[19] Richards, *supra* note 4 at p. 8.
[20] *Ibid.*
[21] *Ibid.*

"In an economy that exploits personal data, the battles over privacy will ultimately determine the allocation of power in our economy and our society as a whole".[22]

Additionally, the idea of "surveillance capitalism" which captures and commodifies personal data to target technology users (ex. companies that monitor internet activities to send targeted advertisements) is an increasingly common occurrence in the digital age.[23] It is clear then, that AI is extenuating the power dynamics between technology creators and individual users. Indicative of this power dynamic is "the way public discussion about AI is influenced by actors developing these technologies".[24] For example, Facebook founder Mark Zuckerberg's 2010 proclamation that privacy is over.[25] As argued by scholars like Richards, Zuckerberg has a clear interest in the dissolution of privacy protection: "many of those calling so loudly for the death of privacy are really seeking its demise so that they can line their own pockets".[26]

The ability to collect and exploit personal information is power, and the consequences of the erasure of individual privacy in this way are severe, resulting in discrimination and marginalization of already-vulnerable groups. For instance, there is a documented practice of employer discrimination by refusing to hire people with disabilities for the express purpose of minimizing healthcare expenditures.[27] Ultimately, when considered against the background of the underlying power dynamics involved, Bill C-27's failure to meaningfully account for individual privacy, leaving critical terms to be defined by the Minister presents ambiguities with serious consequences for individual privacy interests and contributes to technologically produced harms to marginalized populations.

Conclusion

Bill C-27 and the statutes within it ultimately aim to modernize privacy regulation and promote responsible AI creation and innovation in Canada. From a law and economics perspective, Bill C-27 fails to properly balance privacy interests with innovation and development. This short commentary has analyzed Bill C-27 and the intersection of privacy and AI to demonstrate the facade of privacy protection present in contemporary legislation efforts. The findings demonstrate that Bill C-27 overemphasizes innovation and development with serious consequences for individual privacy. Ultimately, if adopted, the *AIDA* will act as a hindrance to individual privacy interests in Canada, reinforcing discriminatory power imbalances and supporting profit over privacy.

[22] *Ibid*.
[23] Matt Malone, A Multidisciplinary Perspective for Policy-Makers, Auditors, and Users, 1st ed (Florida: CRC Press) at p. 73
[24] *Ibid*.
[25] Marshall Kirkpatrick, Facebook's Zuckerberg Says The Age of Privacy is Over (9 January 2010) online: readwrite <https://readwrite.com/facebooks_zuckerberg_says_the_age_of_privacy_is_ov/>
[26] Supra note 22 at 8.
[27] Richards, *supra* note 3 at 8.

References

Jurisprudence

Hunter v Southam, [1984] 2 S.C.R. 145

Government Documents

Bill C-27, An Act to enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act, 1st Sess, 44th Parl, 2022 (at consideration in committee in the House of Commons).

Secondary Sources

Bariffi, Francisco Jose, "Artificial Intelligence, Human Rights and Disability" (2021) 26:2 Pensar - Revista de Ciências Jurídicas 1

Hin-Yan, Liu, "The Power Structure of Artificial Intelligence" (2018) 10:2 Law, Innovation and Technology at 197-229.

Donovan, Joan et al., "Algorithmic Accountability: A Primer" (April 18, 2018) online: Data & society < https://www.datasociety.net/wp-content/uploads/2018/04/Data_Society_Algorithmic_Accountability_Prim- er_FINAL.pdf> at p.11.

Jones, Charles I., and Christopher Tonetti. 2020. "Nonrivalry and the Economics of Data." *American Economic Review*, 110 (9): 2819–58.

Kornhauser, Lewis, "The Economic Analysis of Law", *The Stanford Encyclopedia of Philosophy* (Spring 2022 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2022/entries/legal-econanalysis/>.

Malone, Matt, *A Multidisciplinary Perspective for Policy-Makers, Auditors, and Users*, 1st ed (Florida: CRC Press)

Heikkilä, Melissa, "AI companies are finally being forced to cough up for training data" (July 2 2024) online: MIT Technology Review <https://www.technologyreview.com/2024/07/02/1094508/ai-companies-are-finally-being-forced-to-cough-up-for-training-data/>

Müller, Vincent C., "Ethics of Artificial Intelligence and Robotics", *The Stanford Encyclopedia of Philosophy* (Fall 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/fall2023/entries/ethics-ai/>

Richards, Neil, *Why Privacy Matters* (New York: Oxford Press 2022) at 17-34

Smalley, Suzanne, State privacy laws have been crippled by big tech, new report says (February 1 2024) online: The Record <https://therecord.media/state-privacy-laws-big-tech-lobbying-report>.

Scassa, Teresa, "Regulating AI in Canada: A Critical Look at the Proposed Artificial Intelligence and Data Act" (2023) 101:1 Canadian Bar Review 1.

MacDonnell, Vanessa A., "A Theory of Quasi-Constitutional Legislation", 2016 53-2 *Osgoode Hall Law Journal* 508, 2016 CanLIIDocs 4295

Chapter 3

Exploitation Through Algorithmic Training: The Struggle to Protect Creativity in the AI Era

Charlotte Boyd (She/her)

JD Candidate, Lincoln Alexander School of Law

Abstract

The emergence of Artificial Intelligence (AI) has been one of the most significant innovation our society has seen in recent years. However, AI has raised pressing legal and ethical challenges, particularly in the realm of copyright law. Central to AI's functionality is its algorithmic training process, which often relies on vast datasets, including unlicensed copyrighted works, to produce outputs. Among others, the paper explores the US decision of *Kadrey, et al. v. Meta Platforms, Inc.* where Meta is accused of exploiting authors' copyrighted works without consent in the training of the LLaMA language model. Through the lens of Marxist theories such as class struggle and commodification, this paper critiques how unregulated algorithmic training perpetuates economic disparity by appropriating creators' intellectual labour to generate profit for technology conglomerates. The resulting alienation from their works exacerbates systemic exploitation and undermines creators' incentives to disseminate their works.

Keywords: Intellectual Property, Fair Dealing, Copyright, Artificial Intelligence, Marxism

Introduction

The emergence of Artificial Intelligence (AI) has been one of the most significant innovations to our society in recent years. With AI platforms being introduced daily – ranging from programs that assist with researching complicated topics to those creating lifelike digital art – questions arise about the kind of information AI platforms have access to in order to develop these outputs. At the core of all AI platforms are instructions and rules arising from the information used during the training phase that enable these systems to learn, analyze data, and make decisions.

However, recent lawsuits have spotlighted the legal complexities surrounding algorithmic training. The legal battle has been fierce between corporate giants and unions. In 2023, The New York Times (NYT) advanced a lawsuit against Microsoft and OpenAI, accusing them of copyright infringement due to their unauthorized use of NYT's published works in the training of the algorithms that underscore their AI platforms, resulting in word-for-word reproductions of NYT's articles without proper attribution.[1] This case is only one example of the evolving litigation surrounding the use of unlicensed material in algorithmic training of AI models.

The ongoing discussions, lawsuits and latest decisions emerging from the United States (US) raise the critical question: should copyrighted materials be freely accessible for algorithmic training of AI models? Ultimately, through an analysis of recent decisions in like *Bartz v Anthropic PBC* and *Kadrey v. Meta Platforms, Inc*. this paper argues that the use of unlicensed copyrighted works in the algorithmic training processes of AI models demands stringent regulation, if not prohibition in Canada.

This view is founded on the stark class disparities between AI platform conglomerates and individual creators. Current practices of using unlicensed copyrighted works in algorithm training perpetuate a system of exploitation, where the labour of creators is appropriated without consent or compensation. Intellectual property is effectively seized as a new means of protection, as technology companies continue to extract surplus value by leveraging copyrighted works to develop profit-generating AI models. By addressing this issue, we confront a critical junction in the evolution of digital rights and the future of copyright protection in the age of AI.

Specifically in Canada, this paper succinctly explores the concept of fair dealing under Canadian copyright law, emphasizing its potential role in balancing creators' rights with societal benefits, including innovation. The lack of legal precedent in Canada specific to AI warrant prompt regulatory and legal interventions. Recent successful claim on "fair use" in *Bartz v. Anthropic PBC*[2] and *Kadrey v. Meta Platforms, Inc.*[3] may provide judicial insights on how North American courts would address the AI and copyright legal conundrum.

However, commentaries from Canada have tended to suggest that the decision provides no clear judicial guideline specific to Canada. The Canadian "fair dealing" rule which is based on enumerated grounds, rather than the Common law "fair use" doctrine would suggest that the *Bartz* and *Kadrey* decisions would provide very little judicial guidance for Canadian court.[4] Moreover,

---

[1] *The New York Times Company v. Microsoft Corporation et al.*, No. 1:2023cv11195 - Document 344 (S.D.N.Y. Dec. 27, 2024).
[2] 24-cv-05417 (ND Cal 2025).
[3] 23-cv-03417 (ND Cal 2025).
[4] See for example, Tamara Céline Winegust, AI training copies blessed as "fair use" by U.S. Court – Can a similar path be forged in Canada? (July 18, 2025) online: Smart & Biggar https://www.smartbiggar.ca/insights/publication/ai-training-copies-blessed-as-fair-use-by-u-s-court-can-a-similar-path-be-forged-in-canada; David Yi, William Chalmers and Fiona Sarazin, Two US decisions find that reproducing works to train large language models is fair use – Part 3: Comparing the Anthropic and Meta decisions (July 22 2025) online: Norton Rose Fulbright

despite the 'clarity" emerging from the US, there is still fuzziness surrounding the limits of the fair use defence, i.e. on whether the fair use defence will suffice when training was based on the pirated work of authors, as determined in the *Bartz's* case,[5] and whether evidence of significant market dilution, as suggested in the *Meta* decision, could displace fair use arguments.[6]

In any case, the application of the fair dealing exception in the context of AI training remains a complex legal question, with more legal battles ahead both in Canada and the US.[7] In using a Marxist framework, the paper argues for robust regulatory frameworks, including updates to Canada's Copyright Act, to address the inequities in algorithmic training and protect creators in the AI era. By advocating for stringent regulation or prohibition of unlicensed copyrighted material in AI training, this paper confronts the intersection of digital rights, intellectual property, and the evolution of copyright protection in a rapidly advancing technological landscape.

On AI Copyright Infringement - *Kadrey, et al. v. Meta Platforms, Inc.* (2025)

On July 7th, 2023, three prominent American authors, Richard Kadrey, Sarah Silverman, and Christopher Golden (collectively, the "Plaintiffs"), brought a class action lawsuit against Meta Platforms, Inc (Meta, the "Defendant"). The lawsuit includes all persons or entities in the United States who own a copyright in any work used in training data for LLaMA and alleges that Meta infringed their copyrights by using their works to train the large language model (LLM), LLaMA.[8]

In this submission, the Plaintiffs argue that Meta copied their literary works without prior consent as part of the algorithmic training datasets for LLaMA to derive or intend to derive profits and other benefits from the use of the infringed materials, thus depriving the Plaintiffs of the benefits of their works.[9] In the training process, copious amounts of data were copied, and expressive information was extracted from the copied works.[10] The Plaintiff addressed that their books were specifically part of a sub-dataset called "Books3," which is a part of a larger dataset consisting of fiction and nonfiction books called "The Pile." Meta has admitted to using "The Pile" to train LLaMA.[11] Additionally, "The Pile" contained the entire contents of a website called "Bibliotik," a "shadow library" website which houses an extensive collection of copyrighted materials.[12]

<https://www.nortonrosefulbright.com/en-ca/knowledge/publications/6c3dd9c0/selon-deux-decisions-aux-etats-unis-la-reproduction-d-oeuvres-aux-fins-d-entrainement-de>.

[5] See page 18-19 of Bartz v Anthropic where District Judge William Alsup stated that "*This order doubts that any accused infringer could ever meet its burden of explaining why downloading source copies from pirate sites that it could have purchased or otherwise accessed lawfully was itself reasonably necessary to any subsequent fair use.* There is no decision holding or requiring that pirating a book that could have been bought at a bookstore was reasonably necessary to writing a book review, conducting research on facts in the book, or creating an LLM. *Such piracy of otherwise available copies is inherently, irredeemably infringing even if the pirated copies are immediately used for the transformative use and immediately discarded*. [Emphasis added].

[6] See page 39 of Kadrey v. Meta Platforms, Inc where District Court Judge Vince Chhabria stated that "In cases involving uses like Meta's, *it seems like the plaintiffs will often win*, at least where those cases have better-developed records on the market effects of the defendant's use. *No matter how transformative LLM training may be, it's hard to imagine that it can be fair use to use copyrighted books to develop a tool to make billions or trillions of dollars while enabling the creation of a potentially endless stream of competing works that could significantly harm the market for those books*." [Emphasis added].

[7] On latest lawsuit in Canada, see Statement of Claim in Canadian News Media Companies v OpenAI (November 28, 2024) online: https://litigate.com/assets/uploads/Canadian-News-Media-Companies-v-OpenAI.pdf>; Some Canadian scholars are skeptical of the strength of ongoing case in Canada, See Michael Geist, Canadian Media Companies Target OpenAI in Copyright Lawsuit But Weak Claims Suggest Settlement the Real Goal (December 3, 2024) online: Michael Geist <https://www.michaelgeist.ca/2024/12/canadianmediaopenai/>

[8] Kadrey, R et al v Meta Platforms, Inc 3:23-cv-cv-03417-VC (ND Cal 2023).

[9] *Ibid* at 9.

[10] *Ibid* at 7.

[11] *Ibid* at 5.

[12] *Ibid* at 5.

Nevertheless, Meta trained the algorithms underlying LLaMA with datasets containing the Plaintiffs' copyrighted books, and by design, the training process removed the identifying information about the books' authors and copyright holders.[13] Ultimately, the Plaintiffs contend that the Defendant's use of the Plaintiffs' books in the algorithmic training processes of LLaMA constitutes copyright infringement and violates the *Digital Millennium Copyright Act*. Through this class action, the Plaintiffs seek compensatory remedies for the alleged copyright infringement, as well as injunctive relief requiring Meta to stop using their books to train LLaMA and implement steps to ensure that future outputs of LLaMA do not infringe their copyrights.[14]

The decision in the *Kadrey* case was handed down on the 25th of June 2025 and was a loss for the plaintiff. The decision came two days after the ruling on the *Bartz* case which give considerable weight to the impact GenAI, describing it as "exceedingly transformative" to justify Anthropic's fair use defence – a perspective that favours Big Tech's encroachment on author's intellectual property. The duo decision set the tone for a progressive judicial approach to fair use claim, but critiques have tended to describe it as less of a significant win for technology companies.[15] Commentators have suggested that while the *Bartz* case "pushed the AI copyright debate forward, *Kadrey* hit pause".[16] A deeper review of the case suggest that it is not a strong precedent for Big Tech companies – because the judge indicated that the plaintiff's loss was due in large to the lack of proper records suggesting that significant market harm.[17] However, there is no doubt that this decision places burden on authors by requiring evidence of "market harm", thereby intensifying the struggle between Big Tech and authors.

Marx's Theory and the *Kadrey's* Case as Class Struggle & Exploitation

Marx's theory of class struggle reflects the foundation of capitalism and the root cause of all forms of struggle or conflict within a capitalistic society.[18] The theory of class struggle posits that society is fundamentally divided into two main classes: the bourgeoisie, who own and control the means of production, and the proletariat, who sell their labour.[19] The division between the bourgeoisie and the proletariat creates an inherent conflict of interest, forming the basis of class struggle as demonstrated through the bourgeoisie's exploitation of the proletariat and extracting surplus value from their labour.[20] As workers become aware of their shared interests, they develop class consciousness and engage in collective organization.[21] Marx predicted that increasing class antagonism would eventually lead to the proletarian revolution, overthrowing capitalism and establishing a classless communist society.[22] This theory views class struggle as a primary driver of social evolution, with all of society characterized by conflict between the oppressed and the oppressor.[23]

---

[13] *Ibid* at 7.
[14] *Ibid* at 7.
[15] Keith Kupferschmid, "Kadrey v. Meta Decision: Did Meta Just Win the Battle, But Lose the War?" (June 27, 2025) online: Copyright Alliance <https://copyrightalliance.org/kadrey-v-meta-decision/>.
[16] Whitney Hart, "Kadrey v. Meta: What This Copyright Lawsuit Reveals About the Fragility of AI's Fair Use Defense" (June 28, 2025) online: Driving Influence <https://avenuez.com/blog/kadrey-v-meta-copyright-lawsuit-ai-fair-use-defense/>.
[17] Arthur Gollwitzer, "Federal Courts Find Fair Use in AI Training: Key Takeaways from Kadrey v. Meta and Bartz v. Anthropic" (July 11, 2025) online: Jackson Walker <https://www.jw.com/news/insights-kadrey-meta-bartz-anthropic-ai-copyright/>
[18] Stephen Resnick & Richard D Wolff, "Classes in Marxian Theory" (1981) 13:4 Rev Radical Political Econs 1 at 2.
[19] *Ibid* at 2.
[20] *Ibid* at 2.
[21] *Ibid* at 15.
[22] *Ibid* at 15.
[23] *Ibid* at 15.

The use of unlicensed copyrighted works in training the algorithms that drive AI platforms, such as Meta's LLaMA, leads to the exploitation of the creators and copyright holders. Marx discusses the idea of Marxian exploitation, which is defined as the expropriation of surplus labour such that surplus labour is the excess of time laboured by the worker (the proletariat) over the amount of labour embodied in the bundle of goods the worker consumes, under the assumptions that a worker's whole wage is spent on the bundle.[24] Thus, exploitation occurs due to the amount of dead labour the worker can command through purchasing commodities when his income is less than the amount of labour he expends in production. In other words, Marxian exploitation refers to workers being unfairly compensated for their labour.

In the case of *Kadrey, et al. v. Meta Platforms, Inc.,* Meta holds similarities to the bourgeoisie, as Meta owns and controls the production and output of LLaMA. The creators, including the three named plaintiffs, are comparable to the proletariat, who license access to their copyrighted works and their labour. However, in *Kadrey,* as Meta scrapes the internet for information to use in their datasets, they collect copyrighted works, which they then use in algorithmic training without consent from or compensation for the copyright owners. This process results in the AI platform owners increasing their profits while the copyright owners receive nothing for their labor and ownership. Thus, it can be seen that the technology companies are exploiting the copyright holders and creators. Consequently, such exploitation is what led to the submission of *Kadrey,* where the plaintiff sought remuneration for the damages faced by the creators of the used works.

Unlicensed Alienation from Copyrighted Works: A Glimpse at Marx's Theory of Commodification

The use of unlicensed copyrighted material in training the algorithms that underlie AI platforms, such as LLaMA, alienates creators and copyright holders from their works. By using unlicensed copyrighted works in algorithmic training, technology companies such as Meta are seizing control of creators' and copyright holders' intellectual property to improve their AI technology. The seizing of control of the copyright owner's work creates a division between the creator and their products of labour, which Marx has recognized in his Alienation theory.[25]

Specifically, Marx establishes a labour type of alienation from the product in which the labourer, or in this case, creators and copyright holders, find the works they own and created to be alien to them.[26] Such alienation in an AI era can perpetuate the issues that copyright law seeks to prevent. While the Canadian Copyright Act (the Act) aims to further the public interest by promoting the creation and dissemination of artistic and intellectual works, the unauthorized use of copyrighted materials in AI training raises significant concerns.[27]

This practice has potential to discourage creators from disseminating their works due to the heightened threat of copyright infringement in an environment where algorithmic training remains largely unregulated. In *Kadrey,* the authors have expressed that they know their works are being used and reproduced through Meta's AI platform. They are separated from their creations due to the lack of control over its reproduction and use in AI-generated output, as well as the deliberate removal of the details that identify their authorship.

---

[24] John E Roemer, "Property Relations vs. Surplus Value in Marxian Exploitation" (1982) 11:4 Philosophy & Public Affairs 281 at 281.
[25] Roudro Mukhopadhyay, "Karl Marx's Theory of Alienation" (2020) SSRN Electric Journal.
[26] *Ibid.*
[27] Canadian Intellectual Property Office, "A Guide to Copyright." (23 November 2024), online: <ised-isde.canada.ca/site/canadian-intellectual-property-office/en/guide-copyright>.

While authors, like the three plaintiffs in *Kadrey,* have the legal right to fight the alienation from the works they experience, many marginalized creators and copyright owners do not have the means to pursue such legal action. As a result, they may be unable to continue dissemination their creations, demonstrating Marx's conflict between the upper class and workers. Ultimately, using unlicensed copyrighted materials in the AI training datasets undermine the creative labour market, for which an expansion of the Act that explicitly addresses AI will help mitigate the copyright infringement that occurs in the algorithmic training process.

Canadian Fair Dealings Rule in the Era of AI

The Canadian Copyrights Act aims to balance the rights of creators with the public interest in promoting the creation and dissemination of works.[28] While the Act grants copyright owners exclusive rights over their works, it also provides for exceptions for uses of copyrighted material without infringement, notably the fair dealing exception under s. 29.[29] Specifically, s. 29 recognizes that "fair dealing for the purpose of research, private study, education, parody, or satire does not infringe copyright," which refers to a user's right to utilize copyrighted materials without infringing on the copyright holder's rights.[30]

The Supreme Court of Canada (SCC) has consistently interpreted fair dealing broadly, emphasizing its role in protecting user's rights rather than merely serving as a defence to copyright infringement claims. In the SCC decision of *CCH*, the Court established a two-step test for fair dealing.[31] The defendant must demonstrate that the dealing was for an allowable purpose such as research (which has a broad, liberal interpretation, even including commercial purposes), private study, education, parody, or satire and that the dealing was fair, which is a "question of fact and depends on the facts of each case," requiring a contextual analysis on many different factors (i.e., purpose, character, alternatives, and nature of the dealing).[32] Ultimately, fair dealing is a vital provision within the Act that serves as a mechanism for balancing the rights of copyright owners with broader societal benefits achieved through using copyrighted materials.

In the case of algorithmic training, such processes could be viewed through a liberal interpretation as research or education and thus fall within the balance of promoting innovation while benefiting creators and society at large. However, it is currently unclear whether algorithmic training should be viewed as fair dealing as there has yet to be a Canadian case setting a precedent for these issues. Additionally, precedents from the US arguably provide little guidance for Canadian courts since fair dealing is based on enumerated grounds, in contrast to the US common law fair use jurisprudence. Therefore, given the complexity and novelty of algorithmic learning and the output of AI models, a case-by-case analysis would be necessary to determine whether specific instances of algorithmic training meet the fair dealing criteria.

Are We En-Route for a Marxist-Inspired Copyright Revolution?

While Marxism argues for a global proletarian revolution that would transform our capitalistic society, abolish private property and thus create a global state of communism,[33] I do not believe that the conflict between large technology corporations' algorithmic training of AI programs and the creators and copyright owners, will lead to a communist revolution. As AI platforms continue

[28] Ibid
[29] Copyright Act, RSC 1985, c C-42, s 29.
[30] *Ibid.*
[31] *CCH Canadian Ltd v Law Society of Upper Canada,* 2004 SCC 13 [*CCH*].
[32] *Ibid* at para 52.
[33] Frederick Engels, "Principles of Communism", (October November1847), online: <www.marxists.org/archive/marx/works/1847/11/prin-com.htm>.

to train their algorithms using new datasets, regardless of whether materials are copyrighted, creators and copyright owners will become increasingly frustrated with the lack of respect for their copyrights. They will fight to have AI companies halt using unlicensed works which will amount to more legal cases, and inevitably, federal law will introduce regulations on how copyrighted works must be handled in algorithmic training processes.

In the meantime, the Canadian government must monitor how other countries are designing and implementing AI and algorithmic learning regulations.[34] As there is currently no standard approach to implementing AI, the transformative nature of AI places the same challenges on governments across the globe to find a balance between innovation and protecting copyright law.[35] Such novel challenges have resulted in governments developing legislation "commensurate to the velocity and variety of proliferating" AI technologies, as well as national AI strategies and the creation of voluntary AI guidelines and standards.[36]

Currently, Canada has limited AI regulations in place, with existing frameworks such as Canada's Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems. This code promotes the responsible development and management of advanced generative AI systems through principles such as accountability, fairness, and transparency.[37] Further efforts are needed to translate consultation reports like the recent Consultation on Copyright in the Age of Generative Artificial Intelligence into actionable and enforceable legislations that protect authors.[38]

Conclusion

Examining the challenges posed by AI algorithmic learning through a Marxist lens underscores the stark tension between technology conglomerates and creators. This tension manifests in the alienation and exploitation of creators and copyright holders, as their intellectual labour is appropriated without consent or appropriate remuneration. As the role of AI continues to expand, the urgency for robust regulatory frameworks in Canada becomes undeniable. Binding legislation must strike a delicate balance of fostering innovation while protecting creators' rights. Only through such measures can we address the current system's inequities and safeguard the future of creative labour in this AI era.

[34] IAPP Research and Insights, "Global AI Law and Policy Tracker" (last updated November 2024), online: < https://iapp.org/media/pdf/resource_center/global_ai_law_policy_tracker.pdf> at 2.
[35] *Ibid* at 2.
[36] *Ibid* at 2.
[37] Government of Canada, "Voluntary Code of Conduct for the Responsible Development and Management of Advanced Generative AI Systems" (September 2023), online: <https://ised-isde.canada.ca/site/ised/en/voluntary-code-conduct-responsible-development-and-management-advanced-generative-ai-systems>.
[38] Innovation, Science and Economic Development Canada, Consultation on Copyright in the Age of Generative Artificial Intelligence (2025)

References

Primary Sources

Legislation

Copyright Act, RSC 1985, c C-42, s 29.

Jurisprudence

Bartz v Anthropic 24-cv-05417 (ND Cal 2025).

CCH Canadian Ltd v Law Society of Upper Canada, 2004 SCC 13

Kadrey v. Meta Platforms, Inc23-cv-03417 (ND Cal 2025).

Kadrey, R et al v Meta Platforms, Inc 3:23-cv-cv-03417-VC (ND Cal 2023).

The New York Times Company v. Microsoft Corporation et al., No. 1:2023cv11195 - Document 344 (S.D .N.Y.2024).

Secondary Sources

Journal Article

Mukhopadhyay, Roudro, "Karl Marx's Theory of Alienation" (2020) SSRN Electric Journal.

Resnick, Stephen, & Wolff, Richard D, "Classes in Marxian Theory" (1981) 13:4 Rev Radical Political Econs 1 at 2.

Roemer, John E, "Property Relations vs. Surplus Value in Marxian Exploitation" (1982) 11:4 Philosophy & Public Affairs 281 at 281.

Online: Websites

Canadian Intellectual Property Office, "A Guide to Copyright." (23 November 2024), online: <ised-isde.canada.ca/site/canadian-intellectual-property-office/en/guide-copyright>.

Engels, Frederick, "Principles of Communism", (October November1847), online: <www.marxists.org/archive/marx/works/1847/11/prin-com.htm>.

Geist, Michael, Canadian Media Companies Target OpenAI in Copyright Lawsuit But Weak Claims Suggest Settlement the Real Goal (December 3, 2024) online: Michael Geist <https://www.michaelgeist.ca/2024/12/canadianmediaopenai/>

Gollwitzer, Arthur, "Federal Courts Find Fair Use in AI Training: Key Takeaways from Kadrey v. Meta and Bartz v. Anthropic" (July 11, 2025) online: Jackson Walker <https://www.jw.com/news/insights-kadrey-meta-bartz-anthropic-ai-copyright/>

Government of Canada, "Voluntary Code of Conduct for the Responsible Development and Management of Advanced Generative AI Systems" (September 2023), online: <https://ised-isde.canada.ca/site/ised/en/voluntary-code-conduct-responsible-development-and-management-advanced-generative-ai-systems>.

Hart, Whitney, "Kadrey v. Meta: What This Copyright Lawsuit Reveals About the Fragility of AI's Fair Use Defense" (June 28, 2025) online: Driving Influence <https://avenuez.com/blog/kadrey-v-meta-copyright-lawsuit-ai-fair-use-defense/>.

IAPP Research and Insights, "Global AI Law and Policy Tracker" (last updated November 2024), online: < https://iapp.org/media/pdf/resource_center/global_ai_law_policy_tracker.pdf> at 2.

Keith Kupferschmid, "Kadrey v. Meta Decision: Did Meta Just Win the Battle, But Lose the War?" (June 27, 2025) online: Copyright Alliance <https://copyrightalliance.org/kadrey-v-meta-decision/>.

Statement of Claim in Canadian News Media Companies v OpenAI (November 28, 2024) online: https://litigate.com/assets/uploads/Canadian-News-Media-Companies-v-OpenAI.pdf>

Winegust, Tamara Céline, AI training copies blessed as "fair use" by U.S. Court – Can a similar path be forged in Canada? (July 18, 2025) online: Smart & Biggar https://www.smartbiggar.ca/insights/publication/ai-training-copies-blessed-as-fair-use-by-u-s-court-can-a-similar-path-be-forged-in-canada;

Yi, David, Chalmers, William, and Sarazin, Fiona, Two US decisions find that reproducing works to train large language models is fair use – Part 3: Comparing the Anthropic and Meta decisions (July 22, 2025) online: Norton Rose Fulbright <https://www.nortonrosefulbright.com/en-ca/knowledge/publications/6c3dd9c0/selon-deux-decisions-aux-etats-unis-la-reproduction-d-oeuvres-aux-fins-d-entrainement-de>.

Reports

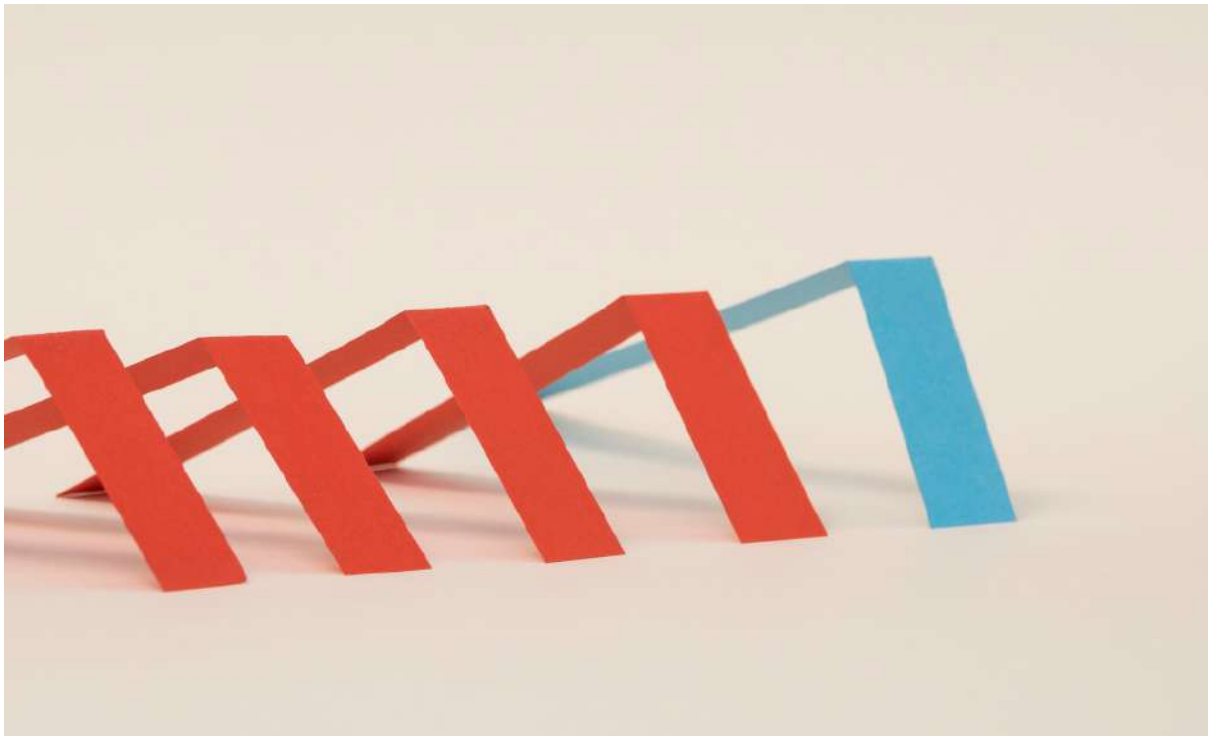
Innovation, Science and Economic Development Canada, Consultation on Copyright in the Age of Generative Artificial Intelligence (2025) .

Chapter 4

Dead End or Hurdle? Unpacking Discrimination in AI-Driven Housing Tools

Amna Sameen Butt

JD Candidate, Lincoln Alexander School of Law



Abstract

The housing industry has been historically encoded with bias and discriminatory practices. With the financial industry now seeking to modernize the housing industry through AI and data-driven tools, the tension between tools that may streamline the process and the risk of further embedding discrimination against persons of colour must be assessed. AI-driven screening and mortgage application tools are used to determine the risk of default of potential tenants and mortgage applicants. This essay will first outline the historical biases that have existed against persons of colour within the housing industry manifesting in the mortgage market and tenant screening. Although AI tools may be advertised as seemingly neutral, I argue that the usage of such tools exacerbates historic biases and works to further marginalize persons of colour and contributes to housing inequality. In doing so this paper uses the critical race theory lens and literature to guide the analysis of how systemic biases are intentionally or unintentionally embedded into algorithms. This essay advocates for the reengineering of algorithms to ensure the anonymity of data to avoid revealing racial markers and embedding bias in outcomes. Further, this essay advocates for reform efforts to allow users on both sides of the process to step in, review the outcome, and correct assumptions made by the technology.

Keywords: artificial intelligence, discrimination, housing industry, bias

Introduction

Discrimination is an undeniable occurrence that has had long and lasting impacts on the housing industry. While discrimination may occur on various grounds, this paper analyzes how discrimination on the grounds of race has been historically embedded into the housing industry and resulted in the further marginalization of racialized individuals.[1] With the increased popularity of AI systems, the housing industry has jumped on the bandwagon of implementing AI in processes such as tenant screening and mortgage applications.[2]

In this paper, I argue that while AI comes with benefits such as that of increased efficiency, the usage of such tool's risks exacerbating historical biases and furthering housing inequality. To mitigate such effects and fix a broken system, this paper advocates for further and ongoing research, the reengineering of algorithms, and greater transparency in AI models and algorithms.

The paper is divided into five sections. Section one is the short introduction above. Section s two presents the historical and ongoing discriminatory practices that are deeply embedded within the housing industry. Section 2? Section three examines how, despite the supposed efficiencies and neutrality of AI, AI-driven housing tools exacerbate biases already rooted within the industry. Section four proposes potential policy and algorithmic solutions that may be used to construct fairer AI-housing tools. Finally, section five provides concluding remarks on the paper.

Past and Present Discrimination in Housing

a. Historical Biases within the Housing Industry

The housing industry has historically been tainted by various discriminatory practices. Unfortunately, these discriminatory practices continue to impact the industry today, resulting in the ongoing disadvantage of already marginalized communities. Discriminatory practices can take various forms but are generally be understood as the intentional or inadvertent, and direct or indirect, unfair and differential treatment of persons based on their personal characteristics.[3] Further, discrimination is not always explicit and may result from policies that are not explicitly prejudicial or from ignorance of existing inequalities.[4]

Discrimination in the housing context is an after-effect of colonial and exclusionary practices against communities deemed undeserving in North America, the effects of which have left or forced these communities into inequitable and unsafe housing.[5] Racialized communities, particularly Black and Indigenous individuals, face high rates of housing insecurity and homelessness due to historical practices such as racial segregation and redlining.[6] For instance,

---

[1] Ontario Human Rights Commission, "V. Identifying discrimination in rental housing" (21 July 2009), online: <www3.ohrc.on.ca/en/policy-human-rights-and-rental-housing/v-identifying-discrimination-rental-housing> [OHRC, "Identifying discrimination in rental housing"].

[2] For an interesting US report by Tech Equity on how AI is utilized in housing screening, see Tech Equity, *Screened Out of Housing: How AI-Powered Tenant Screening Hurts Renters* (July 2024); For a more recent report how AI hikes rent and contribute towards homelessness, see Tech Equity, AI is being used to rent-gouge everyday people: Jesus's story (June 17 2025) online: Tech Equity: <https://techequity.us/2025/06/17/ai-is-being-used-to-rent-gouge-everyday-people-jesuss-story/>; Patrick Sisson, "For Tenants, AI-Powered Screening Can Be a New Barrier to Housing" (September 11 2024) online: Bloomberg <https://www.bloomberg.com/news/features/2024-09-11/ai-powered-tenant-screening-tech-worries-fair-housing-advocates>.

[3] Alexandra Ages et al, *A Path Forward*: *Housing Discrimination in Canada: Urban Centres, Rental Markets and Black Communities* (Ottawa: Max Bell School of Public Policy, 2021) 13.

[4] *Ibid*; Adriano Tesolin, "Anti-Black Racism in Canadian Housing: The Need for Race Based Data" (May 2023), online: <mosaicinstitute.ca/anti-black-racism-in-canadian-housing>.

[5] Centre for Equality Rights in Accommodation, *National Right to Housing Network & Social Rights Advocacy Centre, Housing Discrimination & Spatial Segregation in Canada* (UN Special Rapporteur on Adequate Housing, 2021) 3.

[6] *Ibid*; Ages et al, *supra* note 3 at 25.

redlining involves financial institutions marking racialized communities on maps as zones of poor creditworthiness and subsequently denying loans and mortgages to people living in those zones.[7]

Practices such as these, which denied loans and mortgages to Black and Indigenous communities, displaced them from the rest of society. Individuals were forced into the 'unattractive' outer parts of cities where housing was inadequate.[8] Unfortunately, these communities were not even safe from corporate displacement in the neighbourhoods they came to call home.[9] These underfunded communities, such as Africville, were often bought out by large investors and cities, who industrialized or remodeled these areas, making them unaffordable to its residents.[10] Such practice further displaced racialized individuals, often leaving them homeless.

The practice of displacing racialized individuals from their communities persists today, especially with the rise of the financialization of housing, where housing is bought and sold as a "portfolio asset for speculation".[11] As housing has become more of a commodity rather than a social good, marginalized communities are bought out by big-money investors and remodeled making them unaffordable to its previous residents.[12] These historic challenges and practices have left a lasting impact and continue to be used today.[13]

Racialized communities are found stuck in a vicious cycle, as I discuss next, in which seemingly neutral tenancy and mortgage requirements continue to perpetuate inequalities, create barriers, and discriminate based on race. For instance, the effects of historical discrimination has forced marginalized people to live in redlined areas. Their history of residing in those areas continues to be taken as red flags, leading to housing denial in other spaces despite other positive factors.

b. Current and Persisting Discriminatory Practices

The enduring effects of racially discriminatory practices in housing have reinforced vicious cycles of inequality that limit housing access for persons of colour. Discrimination in housing manifests in various forms, ranging from blatant racism to more subtle and systemic forms of racism.[14] These discriminatory practices may include an outright denial of housing, charging higher prices or rent, and applying more stringent criteria to individuals based on race.[15] For example, three prospective landlords who initially showed interest in a young Jamaican Canadians' housing application

---

[7] Community Housing Transformation Centre, "Reinforcing Black Canadian communities through housing transformation" (13 February 2023), online: <centre.support/reinforcing-black-canadian-communities-through-housing-transformation/>.
[8] *Ibid;* Channon Oyeniran, "Anti-Black Racism in Canada" (1 June 2022), online: <thecanadianencyclopedia.ca/en/article/anti-black-racism-in-canada>.
[9] *Supra* note 3 at 34.
[10] S*upra* note 7 at 34 - 36.
[11] *Supra* note 7 at 23.
[12] *Ibid*.
[13] *Ibid*; Channon, *supra* note 7; Sylvia Novac et al, *Housing Discrimination in Canada: The State of Knowledge* (Ottawa: Canada, Mortgage and Housing Corporation, 2002).
[14] Canadian Centre for Housing Rights, *Housing Equality for New Canadians: Measuring Discrimination in Toronto's Rental Housing Market* (Toronto: Centre for Equality Rights in Accommodation, 2013); The Habitat Group, "Avoid Discrimination When Using AI-based Tenant Screening Services" (24 June 2024), online: < https://www.thehabitatgroup.com/articles/9307-avoid-discrimination-when-using-ai-based-tenant-screening-services>; Channon, *supra* note 8.
[15] Sylvia Novac et al, Housing Discrimination in Canada: What Do We Know About It? (Centre for Urban and Community Studies, December 2002) 1-7; See also Charlton McIlwain, "Can Technology Help Make Housing Fairer?" (2020) 123:6 MIT Technology Rev 44.

denied him when he arrived in person, only to later accept his white friend's application, who, aside from race, had similar characteristics.[16]

Moreover, policies and practices used by financial institutions and landlords – like credit score and reference checks – although seemingly neutral, are often discriminatory and lead to biased outcomes that disproportionately affect racialized individuals.[17] Credit score checks, for instance, are likely to be lower or non-existent for racialized individuals due to a lack of access to equitable credit opportunities.[18] As such, the use of racially coded information, such as credit scores, does not always speak to how good a tenant may be; nevertheless, landlords and mortgage approvers often use these information to justify denying housing.[19] Furthermore, the excessive over-policing of racialized communities can lead to minor criminal records that hinder and individuals' ability to secure housing and do not reflect their true personalities.[20] Thus, policies such as these prevent many racialized individuals from ever having a chance of being "good" tenants or homeowners.[21]

Although various legislations in North America, such as the *Fair Housing Act* in the United States, the *Canadian Charter of Rights and Freedoms*, and human rights legislations within Canadian provinces, enshrine protections against housing discrimination, such laws has been insufficient in remedying barriers and discrimination faced by racialized individuals.[22] In Canada, surveys show widespread discrimination in the rental market. [23] Despite this, housing-related complaints only makeup a small fraction of cases before human rights tribunals.[24] A lack of reporting can have various causes, including difficulty in proving racism due to its systemic nature, limited access to legal representation for racialized individuals, lack of knowledge of one's rights, and a fear of deportation for racialized newcomers.[25] As the housing industry modernizes, blatant, visible, or unspoken biases can be encoded in AI-driven screening and mortgage application tools, to the further detriment of racialized individuals.

The Role of AI in Housing: A Double-Edged Sword

c. Bright Lights

In attempts to modernize the housing industry, financial institutions and landlords have increasingly started adopting AI models to assess credit risk, automate mortgage approvals, and streamline tenant screening processes. For instance, Canada's second-largest lender, Toronto-Dominion Bank, uses generative AI for mortgage pre-approvals.[26] There are benefits to the use of AI within the housing industry. By improving the efficiency with which housing-related

[16] Sameer Chhabra, "This Windsor man took a rental discrimination case to the Ontario Human Rights Tribunal and won" (16 April 2020), online (news): <cbc.ca/news/canada/windsor/windsor-resident-human-rights-tribunal-win-1.5533850>.
[17] Ages et al, *Supra* note 1; Ontario Human Rights Commission, "Housing Discrimination and the Individual" (21 July 2009), online: <www3.ohrc.on.ca/en/right-home-report-consultation-human-rights-and-rental-housing-ontario/housing-discrimination-and> [OHRC, "Housing Discrimination and the Individual"]; The Habitat Group, *supra* note 13.
[18] The Habitat Group, *supra* note 14.
[19] *Ibid; Supra* note 3 at 15; OHRC. *supra* note 17.
[20] Ages et al, *supra* note 1 at 50; Channon, *supra* note 8.
[21] Ages et al, *supra* note 1 at 24; The Habitat Group, *supra* note 14.
[22] Canadian Centre for Housing Rights, *supra* note 14; Centre for Equality Rights in Accommodation, *supra* note 14.
[23] S*upra* note 6 at 10.
[24] *Ibid.*
[25] OHRC, "Housing Discrimination and the Individual", *supra* note 17.
[26] Kelsey Wolfe, "TD using AI to speed up mortgage applications: Artificial Intelligence" (4 June 2024), online: <ezproxy.lib.torontomu.ca/login?url=https://www.proquest.com/newspapers/td-using-ai-speed-up-mortgage-applications/docview/3064163439/se-2?accountid=13631>.

applications are analyzed, including providing quick turnover for applicants, quickly assessing risk, and reducing costs, the transformation of the industry through AI is truly undeniable.[27]

One advantage that creators of AI-driven housing tools boast is their neutrality and power to provide bias-free outcomes unlike those made by human decision-makers.[28] However, despite these benefits and the seeming neutrality of these tools, there is the challenge of AI models perpetuating biases. Biases may be exacerbated through use of historically prejudiced training data and the continuing use of discriminatory requirements, such as credit score analysis, in the decision-making process. Thus, as the widespread need for efficiency increases adoption of AI models in the housing industry, this need must be assessed against the substantial drawbacks of magnifying inequity in an already unequal housing market.

d. Bias in AI Models

AI models are limited to a defined set of data and parameters.[29] This can pose problems such as representation and historical bias due to a lack of neutrality of the training data.[30] There are various ways in which bias can result; for instance, representation bias occurs when a dataset used to train an AI model fails to adequately represent all relevant groups.[31] This issue is likely to be amplified in jurisdictions such as Canada, where a lack of research on housing discrimination can lead to skewed and discriminatory results.

Moreover, the problem of historical bias occurs when AI models reinforce past discriminatory practices, such as redlining.[32] These problems persist even when training data is not overtly racial in nature[33]. For instance, normalized housing practices can reveal racial markers and perpetuate bias, even if the data appear facially neutral. For instance, a check of previous addresses serve as a proxy for racial indicators by revealing whether an individual lived in a low-income community based on postal and ZIP codes.[34] Geographic data, which is highly corelated with race due to historically discriminatory practices, can give away one's racial identity, leading to biased outcomes.[35] This phenomenon is also known as omitted variable bias, which occurs when important factors such as race are left out of the model but are indirectly captured through other variables.[36] Furthermore, discriminatory practices that have become normalized within the housing industry create a harmful feedback loop that can magnify discrimination in an already unequal housing market.[37]

Bias may also occur in the form of aggregation bias, where generalized assumptions about certain groups— for instance, Black individuals being violent or South Asians being smelly — can result

---

[27] *Ibid*; Vanessa G Perry, Kristen Martin & Ann Schnare, "Algorithms for All: Can AI in the Mortgage Market Expand Access to Homeownership?" (2023) 4:4 AI 888 at 889.
[28] Perry, Martin & Schnare, *supra* note 26; Demetria Gallegos, "When AI Denies Your Loan Application, Should You Be Able to Appeal to a Human?" (6 November 2023), online: <wsj.com/tech/ai/ai-denies-loan-application-appeal-to-human-48d18d57#>.
[29] Ibid, Gallegos.
[30] *Ibid*; Dominique Payette & Virgina Torrie, "AI Governance in Canadian Banking: Fairness, Credit Models, and Equality Rights" (2020) 36:1 BFLR 2.
[31] Perry, Martin & Schnare, *supra* note 27; Ana CB Garcia, Marcio GP Garcia & Rigobon Roberto, "Algorithmic Discrimination in the Credit Domain: What Do We Know about It?" (2024) 39: 4 AI & Society 2059.
[32] Sylvia Novac et al, *Housing Discrimination in Canada: The State of Knowledge* (Ottawa: Canada, Mortgage and Housing Corporation, 2002).
[33] The Habitat Group, *supra* note 14; Perry, Martin & Schnare, *supra* note 27; Ages et al, *supra* note 3.
[34] Garcia, Garcia & Roberto, *supra* note 31; "Algorithms for All: Can AI in the Mortgage Market Expand Access to Homeownership?" (2023) 4:4 AI 1.
[35] Dominique & Virgina, *supra* note 30.
[36] "Algorithms for All: Can AI in the Mortgage Market Expand Access to Homeownership?" (2023) 4:4 AI 1
[37] Perry, Martin & Schnare, *supra* note 27.

in biased outcomes in individual cases.[38] Analyzing and mitigating these biases can be difficult, especially since that AI models are typically "black boxes" that has no ability to explain how an outcome is reached.[39] The foregoing problems show how AI tools' intrusive approaches that go beyond traditional checks to arrive at a more 'complete' picture of potential tenants are more problematic than helpful.[40] Unfortunately, due to the novelty of this subject and a lack of academia available, further research must be done to unveil the effects of using AI in order to truly mitigate bias and eliminate inequality in the Canadian housing industry. Regardless, based on existing research, certain solutions may be posed to reconcile the disadvantages of utilizing AI models to arrive at a fairer housing system.

Paving a Path Forward – Policy and Algorithmic Solutions for Reengineering a Fairer Housing System

a. Reflecting on Personal Biases

In retrospect, the research conducted on housing discrimination for this paper has revealed to me personal biases that have unintentionally led me to have likely denied good applicants for my dad's rental properties. An important takeaway from this research is that racialized individuals, such as myself, are not necessarily immune from falling into the shambles of applying discriminatory practices. In fact, with the housing industry having normalized practices that give way to bias and discrimination, it becomes hard for one to reflect on their own biases and their effects. The largest challenge in Canada is the outright ignorance and lack of acknowledgment of racism that exists.[41] As such, not only should more research be done, but research should also be made readily available through public access and education to bring the racist workings behind these processes to the forefront and engender a change that truly tackles them.

b. Auditing and Human Accountability Measures

With the continuous evolution of AI and its fast-paced integration into financial services, the regulatory landscape must take steps to keep pace. Regular monitoring and evaluation of AI systems and the housing industry must be completed to ensure a fair, transparent, and discrimination-free housing industry in Canada.[42] Regular monitoring and evaluation of the housing industry and its related AI systems should be conducted to assess for biases.[43] These audits should focus on data used to train algorithms and factors used by the algorithms to arrive at a final decision. Where audits reveal the use of unnecessary or discriminatory information, systems should be corrected to ensure algorithms are trained to remove or minimize the impact of such information in the final decision.[44] For instance, the use of low credit scores should be eliminated in the final decision-making process if bank statements reveal that a prospective tenant has consistently paid their rent. Furthermore, improving the regulatory landscape also calls for robust regulations that implement accountability measures to ensure developers and financial institutions are held responsible and penalized for any biased outcomes of their AI systems.[45]

---

[38] *Ibid,* OHRC, "Housing Discrimination and the Individual", *supra* note 17.
[39] Dominique & Virgina, *supra* note 30.
[40] Mara Ferreri & Romola Sanyal, "Digital Informalisation: Rental Housing, Platforms, and the Management of Risk" (2022) 37:6 Housing Studies 1035.
[41] *Supra* note 3.
[42] Payette & Torrie, *supra* note 30.
[43] Gary Rhoades, "Ghosts in the Machine: How Past and Present Biases Haunt Algorithmic Tenant Screening Systems" (2024) 49:4 Chicago: American Bar Assoc 13.; Mirka S Caron, "The Transformative Effect of AI on the Banking Industry" (2019) 34:2 BFLR 169; Canadian Centre for Housing Rights, *supra* note 14.
[44] Perry, Martin & Schnare, *supra* note 27.
[45] Demetria Gallegos*; Supra note 18.*

c. Transparency of AI Models and Algorithms

Transparency in AI models and algorithms is necessary to combat discrimination and address concerns of bias in housing.[46] To achieve this, developers of these systems, financial institutions, and landlords should ensure that algorithms used for default risk screening and tenant screening are not black boxes. Transparency of algorithms, AI systems, the types of data used, and efforts undertaken to mitigate bias is essential for ensuring that the decision-making process is understandable and explainable to consumers, affected communities, and regulators alike.[47] Such transparency will not only allow for better and fairer algorithms but will allow systems to continuously improve and help deployers and creators reach the neutrality that they currently wrongly advertise. Clear documentation on how models were trained, the type of data used, and step-by-step explanations of how decisions were made can avoid black box systems and ensure that systems are kept accountable and discriminatory outcomes are mitigated.

d. Anonymization of Data and Quality Checks

Another way to reduce bias in AI models is by reengineering algorithms to ensure the anonymization of data used for training and decision-making.[48] This can involve training algorithms to remove or mask various indicators of race, which may result from, for instance, geographic indicators such as postal codes. Anonymizing data can ensure that AI models rely more on factors that truly reflect an individual's ability to be a good tenant or of a low risk of default, rather than factors that embed historical discriminatory biases. Furthermore, another crucial aspect that can help prevent AI systems from perpetuating bias and discrimination is human oversight as a safeguard.[49] Just as products receive quality checks, creators and deployers of AI models should ensure a human is kept in the loop to quality check final outcomes for discrimination and bias. Human oversight can involve requiring periodic audits, appeal processes, and allowing users on both sides of the process to step in, review outcomes, and correct biased assumptions made by the technology.

Conclusion

Discrimination, particularly on the grounds of race, has been historically embedded in the housing industry through practices that disproportionately affect racialized communities. With the dawn of financial modernization and the introduction of AI, the housing industry risks further exacerbating discrimination and inequities among already segregated and marginalized groups. While AI tools are advertised as neutral, this paper argues that the usage of such tools exacerbates historical biases and works to further housing inequality. To mitigate such effects and fix an already broken system, this paper advocates for further and continuous research, reengineering of algorithms, and transparency of AI models and algorithms.

[46] The Habitat Group, *supra* note 14; Perry, Martin & Schnare, *supra* note 27.
[47] Payette & Torrie, *supra* note 30; Caron, *supra* note 43.
[48] Danya Sherbini, "AI is Making Housing Discrimination Easier than Before" (12 February 2024), online: <kreismaninitiative.uchicago.edu/2024/02/12/ai-is-making-housing-discrimination-easier-than-ever-before/>.
[49] Payette & Torrie, *supra* note 30; Garcia, Garcia & Roberto, *supra* note 31.

References

Journal Articles

Caron S, Mirka, "The Transformative Effect of AI on the Banking Industry" (2019) 34:2 BFLR 169.

Ferreri, Mara & Sanyal, Romola, "Digital Informalisation: Rental Housing, Platforms, and the Management of Risk" (2022) 37:6 Housing Studies 1035.

Garcia GP, Marcio, & Roberto, Rigobon, "Algorithmic Discrimination in the Credit Domain: What Do We Know about It?" (2024) 39: 4 AI & Society 2059.

McIlwain, Charlton, "Can Technology Help Make Housing Fairer?" (2020) 123:6 MIT Technology Rev 44.

Payette, Dominique, & Torrie, Virgina, "AI Governance in Canadian Banking: Fairness, Credit Models, and Equality Rights" (2020) 36:1 BFLR 2.

Perry G, Vanessa, Martin, Kristen & Schnare, Ann, "Algorithms for All: Can AI in the Mortgage Market Expand Access to Homeownership?" (2023) 4:4 AI 888 at 889.

Rhoades, Gary, "Ghosts in the Machine: How Past and Present Biases Haunt Algorithmic Tenant Screening Systems" (2024) 49:4 Chicago: American Bar Assoc 13.

Reports

Ages, Alexandra, et al, *A Path Forward*: *Housing Discrimination in Canada: Urban Centres, Rental Markets and Black Communities* (Ottawa: Max Bell School of Public Policy, 2021) 13.

Canadian Centre for Housing Rights, *Housing Equality for New Canadians: Measuring Discrimination in Toronto's Rental Housing Market* (Toronto: Centre for Equality Rights in Accommodation, 2013).

Centre for Equality Rights in Accommodation, *National Right to Housing Network & Social Rights Advocacy Centre, Housing Discrimination & Spatial Segregation in Canada* (UN Special Rapporteur on Adequate Housing, 2021) 3.

Novac, Sylvia, et al, *Housing Discrimination in Canada: What Do We Know About It?* (Centre for Urban and Community Studies, December 2002) 1-7.

Novac, Sylvia, et al, *Housing Discrimination in Canada: The State of Knowledge* (Ottawa: Canada, Mortgage and Housing Corporation, 2002).

Tech Equity, *Screened Out of Housing: How AI-Powered Tenant Screening Hurts Renters* (July 2024)

Online: Websites

Chhabra, Sameer, "This Windsor man took a rental discrimination case to the Ontario Human Rights Tribunal and won" (16 April 2020), online (news): <cbc.ca/news/canada/windsor/windsor-resident-human-rights-tribunal-win-1.5533850>.

Community Housing Transformation Centre, "Reinforcing Black Canadian communities through housing transformation" (13 February 2023), online: <centre.support/reinforcing-black-canadian-communities-through-housing-transformation/>.

Gallegos, Demetria, "When AI Denies Your Loan Application, Should You Be Able to Appeal to a Human?" (6 November 2023), online: <wsj.com/tech/ai/ai-denies-loan-application-appeal-to-human-48d18d57#>.

Ontario Human Rights Commission, "Housing Discrimination and the Individual" (21 July 2009), online: <www3.ohrc.on.ca/en/right-home-report-consultation-human-rights-and-rental-housing-ontario/housing-discrimination-and>

Ontario Human Rights Commission, "V. Identifying discrimination in rental housing" (21 July 2009), online: <www3.ohrc.on.ca/en/policy-human-rights-and-rental-housing/v-identifying-discrimination-rental-housing>
Oyeniran, Channon, "Anti-Black Racism in Canada" (1 June 2022), online: <thecanadianencyclopedia.ca/en/article/anti-black-racism-in-canada>.

Sherbini, Danya, "AI is Making Housing Discrimination Easier than Before" (12 February 2024), online: <kreismaninitiative.uchicago.edu/2024/02/12/ai-is-making-housing-discrimination-easier-than-ever-before/>.

Sisson, Patrick, "For Tenants, AI-Powered Screening Can Be a New Barrier to Housing" (September 11, 2024) online: Bloomberg
<https://www.bloomberg.com/news/features/2024-09-11/ai-powered-tenant-screening-tech-worries-fair-housing-advocates>.

Tech Equity, AI is being used to rent-gouge everyday people: Jesus's story (June 17, 2025) online: Tech Equity: <https://techequity.us/2025/06/17/ai-is-being-used-to-rent-gouge-everyday-people-jesuss-story/>;
Tesolin, Adriano, "Anti-Black Racism in Canadian Housing: The Need for Race Based Data" (May 2023), online: <mosaicinstitute.ca/anti-black-racism-in-canadian-housing>.

The Habitat Group, "Avoid Discrimination When Using AI-based Tenant Screening Services" (24 June 2024), online: < https://www.thehabitatgroup.com/articles/9307-avoid-discrimination-when-using-ai-based-tenant-screening-services>.

Wolfe, Kelsey, "TD using AI to speed up mortgage applications: Artificial Intelligence" (4 June 2024), online: <ezproxy.lib.torontomu.ca/login?url=https://www.proquest.com/newspapers/td-using-ai-speed-up-mortgage-applications/docview/3064163439/se-2?accountid=13631>.

Chapter 5

Accessible AI Hiring: Transforming Job Recruitment With The Social Model of Disability

Marisa Capano (she/her)

JD Candidate, Lincoln Alexander School of Law

Abstract

In the era of AI-driven hiring, recruitment platforms like LinkedIn are transforming candidate selection through automated recommendation systems. While these AI-powered tools enhance efficiency by analyzing profiles and past employment to match candidates with jobs, they often inadvertently perpetuate barriers for disabled job seekers. By relying on narrow, normative metrics such as continuous employment histories and linear career trajectories, AI systems systematically disadvantage candidates with disabilities, whose unique career paths may be shaped by medical needs or societal barriers rather than by a lack of skill or dedication. This paper uses LinkedIn's recommendation algorithms as a case study to examine how AI-driven hiring tools, grounded in the Medical Model of Disability, reinforce exclusionary practices by treating disability as an individual deficiency. This paper addresses these biases by applying the Social Model of Disability, which reframes disability as a product of societal and environmental barriers, proposing an alternative approach to AI design. It advocates for a paradigm shift in AI recruitment technologies towards removing these societal and environmental barriers rather than penalizing candidates for non-traditional work histories. Additionally, this study calls for legal reforms that mandate accessibility standards within AI hiring systems, requiring that accessibility be integrated into these tools from inception. Reimagining AI with accessibility as a core feature will encourage more inclusive hiring practices, ensuring equitable opportunities for differently-abled job seekers and facilitating the development of diverse and inclusive workplaces.

Keywords: social, data, law, hiring, disability, inclusivity.

Introduction

Artificial intelligence ("AI") has revolutionized hiring processes, with digital platforms like LinkedIn and Google leveraging algorithms to efficiently match candidates with job openings. However, these systems often perpetuate systemic discrimination, particularly against persons with disabilities ("PwDs"). By prioritizing uninterrupted career trajectories and upward mobility, which are traditional markers of employability, AI hiring tools embed historical biases into their algorithms.[1]

As a result, highly qualified PwD candidates are excluded from opportunities. For instance, LinkedIn's job-matching algorithm undervalues candidates with non-linear careers, because it fails to account for systemic barriers or medical needs that disrupt conventional employment paths.[2] The Social Model of Disability provides a theoretical framework for addressing these inequities, shifting the focus from individual impairments to structural barriers.

This paper argues that AI hiring systems perpetuate systemic discrimination against PwDs by embedding outdated employability criteria, neglecting inclusive design principles, and failing to comply with modern legal standards. To eliminate these inequities, the analysis will first examine the systemic barriers in AI hiring practices, then explore algorithmic bias through case studies like LinkedIn's job-matching tool. It will then critique existing legal frameworks and their limitations and finally propose solutions through inclusive design principles and regulatory reforms.

Background: Systemic Barriers In AI Hiring and The Social Model of Disability

The Social Model of Disability (SMD) reframes disability as the result of societal and environmental barriers, such as inaccessible workplaces and biased hiring practices, rather than individual impairments.[3] This perspective challenges AI hiring tools that often reflect the outdated Medical Model of Disability, which treats disability as a personal deficit instead of recognizing systemic barriers that shape PwDs' career paths.[4]

Many AI systems are trained on biased historical data that favor continuous employment and linear career progression. These criteria often exclude PwDs, whose career paths typically reflect external barriers rather than a lack of qualifications.[5] For instance, datasets frequently overrepresent candidates with uninterrupted work histories[6], penalizing PwDs who demonstrate

---

[1] Lauren Weber, "Millions of Résumés Never Make It Past the Bots. One Man Is Trying to Find Out Why" (June 22, 2025) online: The Wall Street Journal https://www.wsj.com/lifestyle/careers/ai-resume-screening-hiring-676a4701; involving a lawsuit by a PWD with a two year employment gap, alleging that Workday, one of the largest purveyors of recruiting software, for discrimination, claiming Workday algorithm screened him out, based on his age, race and disabilities. See Mobley v. Workday, Inc., Case No. 23-CV-770.

[2] Sheridan Wall and Hilke Schellmann, "LinkedIn's job-matching AI was biased. The company's solution? More AI" (June 23, 2021) online: MIT Technology Review <https://www.technologyreview.com/2021/06/23/1026825/linkedin-ai-bias-ziprecruiter-monster-artificial-intelligence/>

[3] Denis Newman-Griffis et al, "Definition drives design: Disability models and mechanisms of bias in AI technologies" (2023) 28:1 First Monday 1-28.

[4] *Ibid*.

[5] Max Langenkamp, Allan Costa & Chris Cheung, "Hiring Fairly in the Age of Algorithms" (2020) Social Science Research Network 1 at 7-8.

[6] Supra note 1. For recent update on the case, see Guy Brenner et al, "AI Bias Lawsuit Against Workday Reaches Next Stage as Court Grants Conditional Certification of ADEA Claim" (June 11 2025) online: Proskauer <https://www.lawandtheworkplace.com/2025/06/ai-bias-lawsuit-against-workday-reaches-next-stage-as-court-grants-conditional-certification-of-adea-claim/>; Ryan Bradshaw, "Biases in AI Recruitment Systems and Their Impact" (August 5 2025) online: Apollo Technical <https://www.apollotechnical.com/biases-in-ai-recruitment-systems-and-their-impact/>

adaptability and resilience through non-traditional trajectories. Research shows unemployment rates for PwDs far exceed those of non-disabled workers; a disparity worsened by hiring tools like LinkedIn's algorithm, which misinterpret career breaks as reduced employability.[7]

Furthermore, CV/Resume screeners have not been trained to recognize their unwanted bias toward marginalized people, and as such, these individuals are consequently less likely to be represented by data.[8] These tools often misinterpret atypical trajectories as signs of reduced employability, embedding systemic discrimination into hiring decisions. Shifting towards a disability justice framework, which emphasizes socio-political and economic relationships shaping disability, is essential to create truly inclusive AI hiring systems.[9]

To address these biases, AI systems must prioritize transparency, such as publishing metrics on dataset composition, and adopt inclusive principles like weighting transferable skills over rigid employment histories.[10] Without these reforms, AI tools will perpetuate structural barriers, excluding qualified candidates and limiting workplace diversity. Inclusive AI design must emphasize fairness-aware ranking systems and involve PwDs in development processes to ensure equitable hiring practices.

Systemic Biases In AI Hiring Tools Exclude PwDs

AI hiring tools perpetuate systemic inequities by prioritizing traditional markers of employability, such as continuous employment and linear career paths. The risk of these hiring tools lies not only in their potential to exclude excellent candidates with non-linear paths, but also in their tendency to actively miss qualified individuals simply because their qualifications do not conform to traditional expectations or norms. In other words, these screeners and algorithms are discriminatory for overlooking crucial candidate factors.

While LinkedIn recently introduced a fairness-aware ranking framework to reduce bias in candidate recommendations, reporting a threefold increase in fairness metrics for gender representation, how its talent ranking systems works to promote fairness remains opaque making it difficult to ascertain how it accommodates the peculiarities of arguably PwD.[11] These entrenched biases stem from training data that replicate historical inequities, under-representing PwDs and amplifying structural barriers. As a result, AI systems systematically exclude qualified candidates whose non-linear career paths reflect external challenges, rather than diminished qualifications.[12] Without rethinking the design principles underpinning these tools, AI risks systematically excluding qualified job seekers who do not align with narrow definitions of employability.[13]

Advocates of these systems may argue that traditional employment markers, like continuous employment, enhance efficiency and predictive accuracy and help recruiters identify candidates

---

[7] Selin E Nugent and Susan Scott-Parker, "Recruitment AI Has A Disability Problem: Anticipating and Mitigating Unfair Automated Hiring Decisions" in Maria Isabel Aldinhas Ferreira, Mohammad Osman Tokhi eds Towards Trustworthy Artificial Intelligent Systems (Switzerland: Springer, 2022) 85-96.
[8] *Ibid* at 88.
[9] , Christo Morr el et al, Towards an Accessible Inclusive Artificial Intelligence (AI2), (Canada: Government of Canada, 2024).
[10] S*upra* note 1.
[11] Maarten Buyl et al, "Tackling Algorithmic Disability Discrimination in the Hiring Process: An Ethical, Legal and Technical Analysis" (2022) In 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22) at 2-4.
[12] S*upra* note 8.
[13] Sahin Cem Geyik et al, "Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search" (2019) KDD '19: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining at 2228.

who demonstrate stability, reliability, and long-term commitment— attributes valued by many employers.[14] While these markers may streamline recruitment, they often overlook the value of diversity. Researchers critique AI's reliance on rigid, classification-based models, which fail to account for the diversity of disabilities and non-linear career trajectories, thereby suggesting a shift toward personalization. Personalization through Neurosymbolic AI models,[15] which prioritize edge cases, (outliers rather than majority population) by being able for instance to capture different forms of disabilities and offering tailored hiring accommodation may better represent PwDs' diverse experiences, although proponents have noted that these models may equally risk broad categorization rather than an assessment that is based on their unique abilities and disabilities.[16]

Furthermore, while efficiency is important, it does not need to come at the expense of diversity. Transparency practices, including regular reports on dataset composition and algorithmic metrics, would further ensure recruiter accountability.[17] Research shows that adaptive algorithms have the potential to balance predictive accuracy with fairness by incorporating broader evaluation criteria, such as transferable skills, project-based experience, and adaptability. By redesigning algorithms to evaluate transferable skills and project-based experience, AI systems can both reduce bias and accurately assess candidate potential.

However, these systemic biases cannot be resolved solely through technical fixes. Legal and regulatory frameworks must play a crucial role in addressing these inequities by mandating transparency, accountability, and inclusivity. Without robust legal standards, biased AI systems will continue to exclude qualified candidates and undermine workplace diversity.

Current Legal Frameworks Fail To Protect PWDs

Existing legal frameworks in Canada fail to adequately protect PwDs from discrimination in AI-driven hiring processes. Canadian legislation, such as the *Accessible Canada Act* and the proposed *Artificial Intelligence and Data Act* (Bill C-27), lacks enforceable mechanisms arising from discrimination resulting specific from AI use in hiring. While Bill C-27 represents a step toward responsible AI design by emphasizing non-discrimination, its early-stage status leaves gaps in protections against systemic biases affecting PwDs.[18]

For instance, these frameworks do not mandate transparency reports or algorithmic audits, preventing regulators and advocacy groups from ensuring compliance with accessibility standards. Without this oversight, AI tools reinforce biases that disproportionately exclude PwDs from employment opportunities. Protecting the sensitive personal data of PwDs is essential to fostering trust in AI-driven solutions, as poorly managed privacy protections can exacerbate inequalities,

[14] See Katherine Weisshaar, From Opt Out to Blocked Out: The Challenges for Labor Market Re-entry after Family Related Employment Lapses (2018) 83:1 American Sociological Review 34. Although the paper focused on employment gap due to family related work leave, the author discussed the Skill Deterioration Theory which suggest that employment gaps can signals skill deterioration to employer, making the potential employee an undesirable candidate. The idea can be transposed to PwDs, who depending on the nature of their disability may also require leave.
[15] Neurosymbolic AI20 is an approach that tries to integrate machine learning approaches with symbolic methods to gain the combined benefits of both approaches (e.g., where large datasets are not available and perhaps where less computing power is available and also to help provide explainable or verifiable AI).
[16] Mike Waid, "AI Data-Driven Personalisation and Disability Inclusion" (2021) 5 Frontiers in AI 1 at 5-6.
[17] S*upra* note 5 at 29.
[18] Bill C-27, An Act to enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act and to make consequential and related amendments to other Acts, 1st Sess, 44th Parl, 2022 (second reading 24 April 2023).

deter the adoption of assistive technologies, and undermine the ethical use of AI, particularly in employment contexts.[19]

Current legislation, such as Canada's *Accessible Canada Act*, fails to mandate the necessary safeguards for these risks, leaving PwDs vulnerable to systemic inequities. This exclusion reduces workplace diversity, stifles innovation, and lowers employee satisfaction.[20] Additionally, this poses a risk for social benefits, as employment among people with disabilities can help lift individuals out of poverty, improve life chances, and benefit society as a whole.[21] AI has the potential to enhance access to rights for PwDs, but without proper regulation, it risks perpetuating discrimination.

However, Accessibility Standards Canada recently published an Employment Standard "CAN/ASC-EN Standard" which the Federal Government also published as draft amendments to the Accessible Canada Regulations under the Accessible Canada Act in December 2024[22] Clause 12.2.3 of the Employment Standard specifically addressed the use of Applicant Tracking Systems and Artificial Intelligence in hiring, and mandated several standards that employee should comply with including that employer produce and make public evidence that they have required suppliers to demonstrate they have taken reasonable steps to make sure their products are not discriminatory to candidates with disabilities, telling candidates whether supplemental assessment tools as part of the Artificial Intelligence screening are accessible to persons with disabilities, taking steps to prioritize the mitigation of Artificial Intelligence bias, even when utilizing third-party tools; and ensuring that there are mechanisms for ongoing analysis of the Artificial Intelligence data and algorithms shared for fairness.[23]

This is a step in the right direction, but it also falls short of robust protection. The standard is voluntary and does not provide for right of private action. Even if PwDs would be able to institute an action, the standard, and possibly the draft regulation does not mandate disclosure of proprietary data, meaning that claimants would face significant hurdle to prove that the AI system discriminated against them.

It could be argue that provincial anti-discrimination laws, such as the *Ontario Human Rights Code*, already addressed bias, making additional regulation unnecessary. It may also be contended that current protections apply to both human and AI decisions and caution that additional regulation could stifle innovation. However, such view overlooks the unique challenges posed by AI. Unlike human decision-makers, algorithms operate on large datasets that often replicate historical inequities, making biased decisions at scale. Adaptive algorithms, designed to account for non-linear trajectories commonly experienced among PwDs, demonstrate that inclusivity and efficiency are not mutually exclusive but can, in fact, enhance each other.[24]

To ensure these algorithms are equitable, adopting rigorous tests and practices is critical. For example, the International Convention on the Rights of Persons with Disabilities (ICRPD)

---

[19] Maram Fahaad Almufareh, et al, "A Conceptual Model for Inclusive Technology: Advancing Disability Inclusion through Artificial Intelligence" 3:1 J Disability Research.
[20] Francisco Jose Bariffi, "Artificial Intelligence, Human Rights and Disability" (2021) 26:2 Pensar - Revista de Ciências Jurídicas 1.
[21] *Supra* note 1.
[22] Canada Gazette, Part I, Volume 158, Number 51: Regulations Amending the Accessible Canada Regulations (December 21, 2024) <https://gazette.gc.ca/rp-pr/p1/2024/2024-12-21/html/reg5-eng.html>
[23] Government of Canada, CAN/ASC-1.1:2024 (REV-2025)-Employment (last visited 4 August 2025) <https://accessible.canada.ca/creating-accessibility-standards/can-asc-112024-rev-2025-employment?mode=full-html> The Standard also included a note mandating that "organization shall ensure that external Artificial Intelligence hiring tools used to conduct screening and evaluation have been programmed and implemented with a data set that includes not only persons with disabilities but diversity within disabilities."
[24] *Supra* note 12.

provides a global framework for regulating AI hiring systems. Through mechanisms like a "digital discrimination test," this framework helps identify inequitable practices and holds systems accountable. However, while the ICRPD is a promising foundation, it faces challenges in capturing the full diversity of disabilities and addressing deeply embedded biases in datasets.[25] These limitations highlight the need for complementary measures, such as participatory design processes and regular algorithmic audits, to fully mitigate risks and promote fairness.

Transparency mandates, routine audits, and accountability standards are vital to ensuring AI systems align with accessibility and fairness principles. Addressing the legal and ethical challenges of AI requires closing the "transparency gap."[26] Moreover, trade secret protections, and limited disclosure often prevent regulators and advocacy groups from assessing algorithmic fairness. As legal protection is not enough, mandatory transparency reports on dataset composition, model metrics, and deployment contexts would provide critical accountability, enabling oversight and fostering public trust.[27] By combining regulation with proactive system design, AI hiring tools can better align with human rights and ethical standards, ensuring equitable opportunities for all candidates. Legal updates should not be viewed as a burden, but as a necessary response to the complexity and harm scalability of AI-driven hiring. While efficiency is essential, it must not come at the expense of diversity.

Inclusive Design is Essential To Transform AI Hiring

Inclusive design practices oriented by SMD are needed to transform AI hiring systems into equitable tools that promote accessibility and diversity. Rather than treating the causes of the challenges that PwDs faces, externalizing it as a social and structural design can better inform empathy and inclusive designs that accommodates the peculiarities and needs of PWDs such as viewing employment gaps and nonlinear career path as a credit. As noted earlier, traditional algorithms undervalue PwD candidates by prioritizing biased criteria that exclude those with non-linear career trajectories shaped by systemic barriers. An SMD inclusive design shifts this focus, encouraging the evaluation of transferable skills, adaptability, and project-based achievements.

AI technologies, such as those developed by Google, Apple, and Microsoft, are promising for empowering PwDs by promoting autonomy and inclusivity in educational and workplace environments.[28] Yet, these advancements must also address financial accessibility and the intersectional needs of individuals with multiple disabilities to achieve true equity.[29] Involving PwDs in AI development ensures that systems address their unique barriers and experiences. Results from the 2022 UN Special Rapporteur's report shares concerns for the risks that AI poses for PwDs, highlighting discriminatory algorithms and biased data sets that can undermine employment opportunities, while also emphasizing the need for PwDs to be actively involved in AI development and policy-making to address these risks effectively.[30] However, it acknowledges

[25] Tetyana (Tanya) Krupiy & Martin Scheinin, "Disability Discrimination in the Digital Realm: How the ICRPD Applies to Artificial Intelligence Decision-Making Processes and Helps in Determining the State of International Human Rights Law" (2023) 23:3 Human Rights L Rev 1 at 23-27.Human Rights Law" (2023) 23:3 HRL Rev at 23-27.
[26] *Supra* note 12 at 5-8.
[27] *Supra* note 5.
[28] Tyler Weitzman, Empowering Individuals With Disabilities Through AI Technology (June 16, 2023) Online: Forbes <https://www.forbes.com/councils/forbesbusinesscouncil/2023/06/16/empowering-individuals-with-disabilities-through-ai-technology/>
[29] Vishal Kumar et al, "The use of artificial intelligence for persons with disability: a bright and promising future ahead" (2024) 19:6 Disability and Rehabilitation: Assistive Technology 1.
[30] *Report of the Special Rapporteur on the Rights of Persons with Disabilities*, UNHRC, 49th Sess, A/HRC/49/52 (2022).

that disability is a fluid and nuanced concept, making it challenging to represent in AI systems trained on static datasets.

Research emphasizes that user-centered design, guided by PwD input, ensures AI tools address real-world barriers effectively.[31] Actively involving PwDs in policy-making processes is essential to ensure generative AI systems prioritize equity, accessibility, and inclusion in both education and broader legal frameworks.[32] Organizations must involve stakeholders, such as HR and IT teams, to evaluate AI tools for fairness and inclusivity before deployment.[33] Transparency measures, including regular audits of datasets and metrics, are vital to identifying and mitigating bias while fostering accountability. While inclusive design principles can mitigate bias in traditional hiring algorithms, the rise of generative AI presents new challenges for ensuring accessibility and engaging with the social model of disability. Generative tools often fail to accommodate diverse disabilities, underscoring the need for policy-making processes that actively involve PwDs.

Some commentators have indicated that inclusive design is costly and *impractical*, which in the context of employment could slow recruitment, and burden companies financially.[34] Additionally, the financial and logistical challenges of redesigning AI systems could slow recruitment processes and place unnecessary burdens on companies, such as small companies with limited resources. While upfront costs may seem high, several scholarly literature have suggested that inclusive design fosters workplace diversity, innovation, and legal compliance, ultimately outweighing initial expenses.[35] This diversity improves the likelihood of employee satisfaction, fosters innovation, and reduces legal risks for organizations.[36] Additionally, involving PwDs in the design process will ensure that AI tools are equitable and effective from the outset, saving time and resources in the long term. Transparent hiring practices and inclusive algorithms not only promote fairness but also enhance the credibility and social responsibility of employers.

As artificial intelligence becomes increasingly embedded in recruitment practices, concerns over bias, accountability, and transparency continue to grow. Policymakers and employers alike face increased demand to ensure ethical standards are upheld. This trend is evident in recent legislative developments in Ontario. Looking ahead, Ontario's *Working for Workers Four Act, 2024* (Bill 149) introduces a notable provision requiring employers to disclose the use of artificial intelligence in external job advertisements, beginning January 1, 2026.[37] This legislative move not only demonstrates the province's commitment to regulating AI in the workplace, but also reflects the growing public and policy concern over opaque algorithmic decision-making in hiring.

Conclusion

AI hiring systems have the potential to transform recruitment, but their reliance on biased criteria perpetuates systemic discrimination against PwDs. Furthermore, adopting the Social Model of Disability is essential to address these inequities. By rethinking algorithmic design, updating legal

[31] Tim Coughlan et al, "Analysing Disability Descriptions and Student Suggestions as a Foundation to Overcome Barriers to Learning" (2024) 2024:1 J Interactive Media in Education.
[32] Katherine C. Aquino et al, "Are institutions addressing accessibility in generative AI policy development?" (2024) 30:2 Disability Compliance Higher Education.
[33] *Supra* note 14.
[34] Marialena Bevilacqua et al, The Return on Investment in AI Ethics: A Holistic Framework (2023) Computers and Society <arXiv:2309.13057> 1>. In commenting on the cost of ethical and inclusive design, the author suggested that commitment to ethical and inclusive AI would only increase if it produces a return on investment.
[35] See for example, Orsa Kekezi, Diversity of experience and labor productivity in creative industries (2021) 55:18 Journal for Labour Market Research 17.
[36] *Supra note 34.*
[37] Ontario, *Bill 149, Working for Workers Four Act, 2024*, 1st Sess, 43rd Parl, 2024.

frameworks, and prioritizing inclusive practices, AI systems can value diversity and equity. Achieving this requires transparency, user-centered design, and stronger legal protections aligned with accessibility standards. Policymakers must enforce algorithmic accountability, and ensure fairness is central to AI's development. A social disability justice lens must address these disparities, ensuring AI's benefits are equitably distributed, while mitigating its costs for vulnerable populations. Future research and international collaboration are crucial to developing frameworks that amplify AI's positive impacts and ensure its equitable application worldwide. By adopting these measures, AI can enrich workplaces with the resilience, adaptability, and innovation that PwDs bring to the workforce.

References

Jurisprudence

Mobley v. Workday, Inc., Case No. 23-CV-770

Legislation

Accessible Canada Act 2019

Bill C-27, An Act to enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act and to make consequential and related amendments to other Acts, 1st Sess, 44th Parl, 2022 (second reading 24 April 2023).

Canada Gazette, Part I, Volume 158, Number 51: Regulations Amending the Accessible Canada Regulations (December 21, 2024)

Ontario, *Bill 149, Working for Workers Four Act, 2024*, 1st Sess, 43rd Parl, 2024.

Journals Articles

Aquino C., Katherine, et al, "Are institutions addressing accessibility in generative AI policy development?" (2024) 30:2 Disability Compliance Higher Education.

Bevilacqua, Marialena, et al, The Return on Investment in AI Ethics: A Holistic Framework (2023) Computers and Society <arXiv:2309.13057> 1>.

Buyl, Maarten, et al, "Tackling Algorithmic Disability Discrimination in the Hiring Process: An Ethical, Legal and Technical Analysis" (2022) In 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22) at 2-4.

Cem Geyik, Sahin, et al, "Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search" (2019) KDD '19: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining at 2228.

Coughlan, Tim, et al, "Analysing Disability Descriptions and Student Suggestions as a Foundation to Overcome Barriers to Learning" (2024) 2024:1 J Interactive Media in Education.

Fahaad Almufareh, Maram, et al, "A Conceptual Model for Inclusive Technology: Advancing Disability Inclusion through Artificial Intelligence" 3:1 J Disability Research.

Jose Bariffi, Francisco, "Artificial Intelligence, Human Rights and Disability" (2021) 26:2 Pensar - Revista de Ciências Jurídicas 1.

Kekezi, Orsa, Diversity of experience and labor productivity in creative industries (2021) 55:18 Journal for Labour Market Research 17.

Krupiy (Tanya), Tetyana, and Martin Scheinin, "Disability Discrimination in the Digital Realm: How the ICRPD Applies to Artificial Intelligence Decision-Making Processes and Helps in Determining the State of International Human Rights Law" (2023) 23:3 HRL Rev at 23-27.

Kumar, Vishal, et al, "The use of artificial intelligence for persons with disability: a bright and promising future ahead" (2024) 19:6 Disability and Rehabilitation: Assistive Technology 1.

Langenkamp, Max, et al, "Hiring Fairly in the Age of Algorithms" (2021) SSRN at 7-8.

Langenkamp, Max, Costa, Allan & Cheung, Chris, "Hiring Fairly in the Age of Algorithms" (2020) Social Science Research Network 1 at 7-8

Newman-Griffis, Denis, et al, "Definition drives design: Disability models and mechanisms of bias in AI technologies" (2023) 28:1 First Monday 1-28.

Mike Waid, "AI Data-Driven Personalisation and Disability Inclusion" (2021) 5 Frontiers in AI 1 at 5-6.

Weisshaar, Katherine, From Opt Out to Blocked Out: The Challenges for Labor Market Re-entry after Family Related Employment Lapses (2018) 83:1 American Sociological Review 34.

Book Chapters

Nugent, Selin E and Scott-Parker, Susan, "Recruitment AI Has A Disability Problem: Anticipating and Mitigating Unfair Automated Hiring Decisions" in Maria Isabel Aldinhas Ferreira, Mohammad Osman Tokhi eds Towards Trustworthy Artificial Intelligent Systems (Switzerland: Springer, 2022) 85-96.

Online: Websites

Bradshaw, Ryan, "Biases in AI Recruitment Systems and Their Impact" (August 5 2025) online: Apollo Technical https://www.apollotechnical.com/biases-in-ai-recruitment-systems-and-their-impact/

Brenner, Guy, et al, "AI Bias Lawsuit Against Workday Reaches Next Stage as Court Grants Conditional Certification of ADEA Claim" (June 11 2025) online: Proskauer https://www.lawandtheworkplace.com/2025/06/ai-bias-lawsuit-against-workday-reaches-next-stage-as-court-grants-conditional-certification-of-adea-claim/

Government of Canada, CAN/ASC-1.1:2024 (REV-2025)-Employment (last visited 4 August 2025) https://accessible.canada.ca/creating-accessibility-standards/can-asc-112024-rev-2025-employment?mode=full-html

Wall, Sheridan, and Schellmann, Hilke, "LinkedIn's job-matching AI was biased. The company's solution? More AI" (June 23, 2021) online: MIT Technology Review https://www.technologyreview.com/2021/06/23/1026825/linkedin-ai-bias-ziprecruiter-monster-artificial-intelligence/

Weber, Lauren, "Millions of Résumés Never Make It Past the Bots. One Man Is Trying to Find Out Why" (June 22, 2025) online: The Wall Street Journal https://www.wsj.com/lifestyle/careers/ai-resume-screening-hiring-676a4701.

Weitzman, Tyler, Empowering Individuals With Disabilities Through AI Technology (June 16 2023) Online: Forbes https://www.forbes.com/councils/forbesbusinesscouncil/2023/06/16/empowering-individuals-with-disabilities-through-ai-technology/

Reports

Morr, Christo, ell et al, Towards an Accessible Inclusive Artificial Intelligence (AI2), (Canada: Government of Canada, 2024).

Report of the Special Rapporteur on the Rights of Persons with Disabilities, UNHCR, 49th Sess, A/HRC/49/52 (2022).

Chapter 6

The Dangers of AI Surveillance in Education: Where Do We Draw the Line?

Lawrence Elkhinovich (He/Him)

JD Candidate, Lincoln Alexander School of Law

Abstract

With the rapid growth and dependence on Artificial Intelligence (AI), educational environments have begun discussing how to integrate this software into their classrooms. These AI tools consist of facial recognition, behavioural pattern trackers, and content monitoring, which are said to improve student safety, security, and emotional well-being. While AI technologies have gained widespread awareness and are employed in virtually every industry and setting, when is it too much? I argue that AI surveillance in educational settings contradicts the principles of fairness and impartiality that schools are meant to uphold. Moreover, AI technologies used in educational environments violate students' privacy rights, particularly given the concern that these students are often minors. AI tools have been known to exacerbate preexisting prejudices, exhibit flaws, and are often linked to pseudoscience. As things stand, AI technology should not be replacing human emotions, and with inherent biases and uncertainties, how can we trust it in schools? This short paper reflects on the use of AI in educational settings and demonstrates why the use of AI in schools is fundamentally dangerous and should not be utilized as a surveillance tool in these environments. By extension, the extensive use of AI surveillance in schools may create a concerning precedent for the future of privacy and autonomy in education if it's implementation is not halted.

Keywords: Student Privacy, Educational Equity, Institutional Trust, and Consent

Introduction

Students are frequently instructed not to utilize artificial intelligence ("AI") to author their assessments. Prior to submitting their work, students are often met with a message confirming that they adhere to the academic misconduct policy, prohibiting the submission of AI-generated content as their own. While some students are given permission by their instructors to use AI for grammar or idea generation, there are a number of strict limitations and warnings associated with this approval. This begs the question, why are educational institutions now seeking methods to incorporate AI as a tool of surveillance into their classrooms? Notably, these technologies are believed to be capable of detecting behavioural trends, monitoring user content, and identifying biometric features.[1]

Utilizing these systems in an educational environment as a tool of surveillance is projected to reduce the workload of teachers and improve overall classroom efficiency.[2] As this concept is still being discussed in many school boards around the nation, now is the time to speak up. In this short piece, I reflect on the use of AI in educational settings. I note that using AI surveillance technologies in the classroom or an academic setting is inappropriate and contradicts the ethos of an equitable and impartial learning environment. With AI use being strongly discouraged in most schools for student use, apart from a few restrictions, educators should not be permitted to use it to encroach on student privacy. The double standard is evident, particularly in an atmosphere where trust, innovation, and student liberties are expected to be respected and upheld.

The Faultiness and Inherent Bias of AI

AI systems are largely built on data, and they often display outputs as a result of embedded biased data.[3] While many claim that machine learning-based surveillance tools are normally impartial, many instances have shown that these systems have perpetuated inherent human biases. For example, Amazon created an AI recruitment tool and programmed it to exclude generic gendered phrases, yet the results still remained skewed.[4] Regardless of how the system was programmed, it continued to perpetuate and use biased gendered data that was already in existence.[5] Moreover, representation bias has shown that AI systems often favour Caucasian individuals, reinforcing systemic racism.[6] Due to the lack of diversity in most data sets, AI systems conventionally perform better when faced with individuals who have a lighter complexion.[7] This results in the targeting of minorities, created by a technology that stems from a lack of equitable data.

With all of this information in mind, AI systems cannot be trusted and relied on for information in an educational setting. How can one be sure that it will remain neutral when it has historically demonstrated contrary tendencies? This is a high risk for schools, particularly for minority and marginalized students, because they might be flagged more frequently and face harsher consequences due to a biased AI system. Counter-proponents may argue that AI tools although not entirely impartial, may lead to a safer and more productive environment. Further, these individuals may contend that human decision-making and judgement is often skewed, so an AI system should also be able to exhibit bias. However, an AI system is not a real person, so it cannot

[1] Daniel, Buck, "AI is a Serious Threat to Student Privacy" (last visited 17 March 2025), online: *Thomas B. Fordham Institute* <>.
[2] Ibid.
[3] Lorenzo Belenguer, "AI Bias: Exploring Discriminatory Algorithmic Decision-Making Models and the Application of Possible Machine-Centric Solutions Adapted from the Pharmaceutical Industry" (2022) 2:771 *AI & Ethics* 771 at 774.
[4] Ibid At 777.
[5] Ibid.
[6] Ibid at 773.
[7] Ibid.

be held responsible for its actions. On the other hand, humans can be held accountable for their actions, which may ultimately act as a deterrent measure. In theory, while an AI surveillance system may make an environment safer, the more important question is who is it making it safer for?

Learning Amid Online Surveillance

Surveillance can't be avoided completely. However, constant content monitoring has been proven to impede student growth.[8] Students require privacy to explore their hobbies, establish new friends online, and mature as young adults, which cannot be accomplished by continuous content monitoring.[9] Students are especially prone to feeling scrutinized, and it can be quite daunting to know that someone is always watching them.[10] As a result, students are not able to completely express themselves, negatively affecting their learning and growth in the long run.[11]

Furthermore, continuous online surveillance may undermine the trust and confidence a student has in their educators.[12] This creates an innate tension, creating greater room for secrecy as the trust is broken.[13] No student would feel comfortable knowing that their instructor or the administration is secretly keeping a closer eye on their search history after revealing something personal to them. In addition, utilizing monitoring to limit expression becomes a technique of control in an environment where students are told to express themselves freely. While content monitoring already exists in school settings to a degree, school boards are contemplating introducing AI programs to further monitor students. This approach seems excessive and burdensome. While some may argue that this approach is advantageous as it minimizes the dangers that youth may face online, a school setting is a place for growth and stimulation. Students should feel like they are trusted to make the right choices.

Examining Safety in Physical Surveillance?

To take this discussion a step further, a report published by the National Association of School Psychologists to determine how students felt when more physical security measures were implemented in their schools showed that security guards, cameras, and metal detectors have raised concerns among students and are said to have a detrimental impact on perceptions of safety.[14] Moreover, as previously stated surveillance technologies, whether virtual or physical, have been shown to negatively impact the overall social climate in a school and function as a tool of creative suppression.[15]

The evidence presented above demonstrates that surveillance across schools is already creating widespread concern among students. However, if schools begin implementing AI surveillance mechanisms, these tools will most likely be even more intrusive and overbearing. AI is likely far more advanced than the tools and systems that schools currently use for surveillance. As such, monitoring will seemingly intensify and as evidence has shown student creativity and exploration will decrease.

---

[8] Danielle Keats Citron, "The Surveilled Student" (2024) 76:1439 *Stanford Law Review* 1439 at 1457.
[9] Ibid.
[10] Ibid.
[11] Ibid.
[12] Ibid at 1458.
[13] Ibid.
[14] National Association of School Psychologists, *School Security Measures and Their Impact on Students* (2018), Bethesda, MD: National Association of School Psychologists at 2.
[15] Ibid at 2-3.

The Panopticon Effect

The notion of the panopticon was initially proposed by Jeremy Bentham and then further elaborated on by Michel Foucault.[16] It serves as an illusion of power and surveillance,[17] pertaining to this discussion. The panopticon is a circular prison structure with a guard watch tower in the middle.[18] All of the prison inmates can be seen and monitored from this watchtower, but they are unable to determine if anyone is occupying the watchtower at any given moment.[19] The idea behind this design was to emphasize that surveillance is constant, and it would be impossible to determine if someone is actually surveilling at that specific moment.[20]

This example demonstrates how power and social control operate, causing inmates to self-discipline their behaviour as they never know when someone is watching.[21] This example extends far beyond a prison and the underlying concept may be utilized in a variety of examples.[22] The panopticon is very relatable to the idea of employing AI to continually monitor and survey students. The same underlying notions of power, control, and discipline are present in both examples, leading to a fear of being watched. As outrageous as it may appear to equate an educational institution to a prison, AI surveillance is making this analogy increasingly relevant.

Conclusion

The use of AI monitoring tools in educational settings presents a significant risk to privacy. While some may argue that AI monitoring allows students to be observed under close supervision, enhancing safety and security, this is far from true. This short reflection has demonstrated that existing educational surveillance already negatively alters the climate of an educational setting. AI surveillance will further exacerbate the risks and barriers present in educational spaces. With AI taking over almost every industry and space, this does not imply that we need to see it implemented everywhere, especially in environments that are already secure and surveilled.

[16] Gilbert Caluya, "The Post-Panoptic Society? Reassessing Foucault in Surveillance Studies" (2010) 16:5 *Social Identities* 621 at 622.
[17] Ibid.
[18] Ibid.
[19] Ibid.
[20] Ibid at 625.
[21] Ibid.
[22] Ibid.

References

Secondary Materials: Articles

Belenguer Lorenzo, "AI Bias: Exploring Discriminatory Algorithmic Decision-Making Models and the Application of Possible Machine-Centric Solutions Adapted from the Pharmaceutical Industry" (2022) 2:771 *AI & Ethics* 771.

Caluya Gilbert "The Post-Panoptic Society? Reassessing Foucault in Surveillance Studies" (2010) 16:5 *Social Identities* 621.

Citron Danielle Keats, "The Surveilled Student" (2024) 76:1439 *Stanford Law Review* 1439.

Secondary Materials: Research Study & Opinion Piece

Buck, Daniel, "AI is a Serious Threat to Student Privacy" (5 October 2023), online: *Thomas B. Fordham Institute* https://fordhaminstitute.org/national/commentary/ai-serious-threat-student-privacy.

National Association of School Psychologists, *School Security Measures and Their Impact on Students* (2018), Bethesda, MD: National Association of School Psychologists.

Chapter 7

Algorithmic Reparation in the Criminal Justice System: Addressing Racial and Gender Bias, Stereotypes and Structural Inequalities within Data-Driven Decision Making

Rajvir Gill (Rav) (She/Her)

JD Candidate, Lincoln Alexander School of Law

Abstract

Algorithms and data-driven technologies are increasingly being used in the criminal justice system to assist in predictive policing, sentencing, parole eligibility, and the risk assessments to predict recidivism. While these technologies have promised efficiency and objectivity, they face criticism for perpetuating racial and gender biases, reinforcing harmful stereotypes, and exacerbating structural inequalities. This short paper argues the need for an algorithmic reparation to address harmful impacts of biased technologies on marginalized communities within the criminal justice system. AI tools can result in disproportionately negative outcomes for racialized communities, which often reflect historical discrimination rather than objective evaluations. Using historical data, such as records from discriminatory policing practices, often results in labeling racialized communities as higher risk to committing crimes leading to negative outcomes. The paper challenges that existing criticisms are not simply a misunderstanding of algorithms, instead they are based on the harm that particularly data-driven technologies cause to marginalized groups, especially Black and Indigenous populations. Drawing eclectically upon critical frameworks, such as race critical code studies, digital caste systems, and surface-level data, this paper argues that AI tools reflect and magnify structural inequalities within the criminal justice system. By examining the intersections of technology, race, gender, the paper emphasizes the need of creating a more inclusive and just legal system rather than undermining the efforts towards equity in the criminal justice system.

Keywords: Algorithmic reparation, Racial bias, Criminal justice system, Predictive policing

Introduction

Algorithms and data-driven tools are being gradually adopted within the criminal justice system, specifically towards technology-driven decision-making in areas traditionally reliant on human judgment.[1] Artificial intelligence is used widely in predictive policing that forecasts potential high-risk neighbourhoods needed for extra patrolling.[2] Algorithms and data-driven tools are often thought to be more fair and less influenced by personal opinions because they use large amounts of data to make decisions instead of relying on human judgment.[3] However, this perception of neutrality is very challenging, as research reveals that algorithms do not operate in isolation from the societal contexts in which they are developed and installed.[4] Algorithms and data-driven tools are instead seen perpetuating systemic inequalities, which are rooted from historical biases.[5]

Data within the criminal justice system is formed from past practices, prejudices, specifically in settings where racialized communities have been over-policed and over-criminalized.[6] To exemplify, data sets that are used in predictive policing may reflect decades of disproportionate policing of Black and Indigenous neighborhoods, rather than an unbiased record of criminal activity.[7] Algorithms that are trained on data that reflects historical biases, such as racially skewed arrest records, will potentially reproduce such biases within their predictions.[8] The problem lies in the idea that the data used to train these algorithms often contains the same biases and unfairness found in human decisions.[9] It is difficult to understand decisions made with algorithms and data-driven tools in its entirety, as we are unable to see the whole process of how such decisions were extracted.[10] This makes it unjust to use such technology in decision making when it can potentially affect people's lives.

The problem of bias in algorithms and data-driven tools goes beyond race, gender biases also influence these tools and often intersects with racial discrimination.[11] It can be argued that these issues are stemmed from the biases that are embedded within the training data and design. It is crucial to further address algorithmic biases within the criminal justice system, as these tools risk exacerbating existing disparities and undermining efforts to promote fairness and equity. This paper explores these challenges through critical frameworks, such as race critical code studies, digital caste systems, and surface-level data emphasizing the need for algorithmic reparation through a legal framework to address these harms. The short paper examines how algorithms reinforce structural inequalities and proposes actionable solutions through algorithmic reparation.

Historical Biased Data

The reliance on historical data in the criminal justice system for the purpose of training algorithms is a critical factor contributing to the perpetuation of systemic biases. Scholars like Gideon Christian has emphasized that AI tools such as risk assessments are often trained on data that reflects historical inequalities, such as biased policing, racial profiling, and discriminatory

[1] Mirko Bagaric et al. "The Solution to the Pervasive Bias and Discrimination in the Criminal Justice System: Transparent and Fair Artificial Intelligence." (2022) 59:95 American Criminal Law Review 95. [Bagaric]
[2] *Ibid* at 110
[3] *Ibid* at 144
[4] Gideon Christian, "Legal Framework For The Use Of Artificial Intelligence (AI) Technology In The Canadian Criminal Justice System" (2024) 21:2 Canadian Journal of law and Technology 109. [Christian]
[5] *Ibid* at 113
[6] *Ibid* at 117
[7] *Ibid* at 117
[8] Aziz Z Huq, "Racial Equity in Algorithmic Criminal Justice" (2019) 68:6 Duke Law Journal 1043. [Huq]
[9] *Ibid* at 47
[10] Christian, *supra* note 4 at 134.
[11] Bagaric, *supra* note 1 at 98

sentencing practices.[12] Through Gideon Christian's analysis it can be established that these tools replicate existing inequalities because the data they rely on is fundamentally biased.

The Supreme Court of Canada (SCC) addressed a similar concern in *Ewert v Canada*, where it ruled that risk assessment tools used for Indigenous offenders were not properly tested, which led to unfair and discriminatory outcomes.[13] Jeffrey Ewert, a Métis individual serving a life sentence, challenged the Correctional Service of Canada's use of risk assessment tools, which had been developed using predominantly non-Indigenous populations.[14] The Court discovered that the Correctional Service of Canada (CSC) had failed to ensure the accuracy of these tools when applied to Indigenous offenders, which demonstrates how reliance on biased data can perpetuate systemic discrimination within the criminal justice system.[15] The SCC stressed that these algorithmic tools must consider special circumstances of Indigenous people, instead of treating everyone the same.[16] This case stresses the idea that algorithmic tools that are developed using data from one demographic group may produce biased outcome when applied to another.

The historical biases are increased when algorithms are treated as impartial decision-making tools. Ruha Benjamin's idea of the "New Jim Code" shows how algorithms, while appealing to be neutral, they strengthen existing racial inequalities.[17] The idea of the "New Jim Code" elaborates on the importance of critically examining these datasets, as they are recognized not to be neutral and are shaped by the same structures of inequality that have historically marginalized racial groups.[18] Additionally, in the case *R v Le* (2019 SCC 34), the Supreme Court highlighted the disproportionate policing of racialized communities, a practice that generates biased datasets.[19] Such practices create a feedback circle where over-policing generates data, which in return updates algorithms that perpetuate these inequities.[20]

In Canada, practices such as carding unfairly target Black and Indigenous people, leading to data that strengthens existing inequalities.[21] To exemplify, data is being used from decade old carding practices that were utilized by Toronto Police, where young African Canadian and Indigenous men were excessively targeted.[22] Furthermore, the "black box" nature of algorithms further sheds light on the issue of historical bias, as decisions are made without full transparency of how the data was collected and applied to the artificial intelligence tools.[23] There is a need to decode the "black box" in algorithmic decision-making, as it is critical to ensure there is transparency and explainability within the criminal justice system.

The *US Loomis* case further demonstrates how the "black box" nature of algorithms can challenge the transparency within legal proceedings within the criminal justice system.[24] The accused within this case tried challenging the risk of recidivism algorithm, COMPAS, however the court denied the accused's ability to challenge the tool's conclusions.[25] Since the accused could not see how the

[12] Christian, *supra* note 4 at 114.
[13] *Ewert* v *Canada*, 2018 SCC 30, 2 SCR 165. [*Ewert*]
[14] *Ibid* at para 77
[15] *Ibid* at para 117
[16] *Ibid* at para 123
[17] Ruha Benjamin, *Race After Technology: Abolitionist Tools for the New Jim Code* (Cambridge, UK: Polity Press, 2019) at 3 [Benjamin]
[18] *Ibid* at 4
[19] *R* v *Le*, 2019 SCC 34, 2 SCR 692 at para 97. [*Le*]
[20] Huq, *supra* note 8 at 27
[21] Christian, *supra* note 4 at 117.
[22] *Ibid* at 118
[23] *Ibid* at 124.
[24] *Ibid* at 125.
[25] *Ibid* at 125.

COMPAS tool generated its decisions, it raised concerns about fairness and transparency within the trial.

Overall, historical biased data is significant in understanding how algorithms perpetuate systemic inequalities within the criminal justice system. When algorithms and data-driven tools rely upon practices such as carding and discriminatory policing, they reinforce racial biases.

Intersectionality: Algorithmic Bias Beyond a Single Category

Examining algorithmic bias through an intersectional lens showcases how data-driven tools can be a disadvantage for those individuals who are dealing with more than one type of discrimination. Kimberlé Crenshaw, an advocate and a scholar of critical race theory, stresses the idea that the failure to address the intersectional impacts of discriminations sustains systemic inequalities.[26] This critical lens is important to have when evaluating technologies like facial recognition and risk assessment tools, which often fail to recognize how intersecting identities have an impact on people's lives.[27] It can be understood that the failure to employ an intersectional lens within algorithmic tools can mask the way algorithms spread existing biases. Facial recognition tools are a prime example of how algorithms fail to address intersectional biases, as these tools are less accurate when identifying people of colour compared to non-coloured.[28] When such algorithms display inaccuracies leading to wrongful convictions, they deepen the existing patterns of discrimination.

According to the 2021 report by the Office of the Correctional Investigator, Indigenous women account for approximately 50% of the federal women's prison population, despite making up less than 5% of the Canadian population.[29] This is due to the risk assessment tools that exacerbate such overrepresentation by excessively labeling Indigenous women as high-risk within the criminal justice system.[30] These risk assessment tools rely upon factors such as employment status, housing stability, and prior interactions with the criminal justice system, which are fixated within the systemic marginalization of indigenous communities.[31] In *R v Gladue* (1999 SCC 1), the SCC affirmed that there is a need to consider the systemic factors contributing to Indigenous overrepresentation in sentencing decisions.[32] However, these factors are not critically examined and evaluated when creating algorithmic tools for risk assessments.[33] By failing to account for intersecting factors, algorithmic tools do not follow the objectives of Gladue principles, which highlight the need for equity in sentencing. In summary, there is a significant need in designing algorithms and data driven tools that incorporate intersectional data in further reducing systemic biases.

Algorithmic Reparation: A Legislative Framework for Equity in the Criminal Justice System

In efforts of addressing systemic harms caused by biased algorithms and data-driven tools, a comprehensive approach based on equity and social justice is essential. Algorithmic reparation goes beyond the simple idea of technological tweaks, it is the need of a legislative reform to ensure accountability, transparency, and inclusivity within the training and application process. Having

---

[26] Kimberle Crenshaw, "Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics" (1989) 139:1 U Chi Legal F 139. [Crenshaw]
[27] Jenny Davis et al, "Algorithmic Reparation" (2021) 8:1 Big Data & Society 1 at 2 [Davis]
[28] Huq, *supra* note 8 at 28
[29] Canada, Office of the Correctional Investigator, *Annual Report 2021-2022* (Ottawa: Office of the Correctional Investigator, 2021) at 96 [Correctional Investigator]
[30] *Ibid* at 25
[31] *Ibid* at 22
[32] *R v Gladue*, 1 SCR 688 at para 69 [*Gladue*]
[33] Correctional Investigator, *supra* note 29 at 98

a legislative change can assist in providing clear guidelines for algorithmic governance and mandating bias and ethical standards.[34] In efforts to embed these principles into the law, it can be ensured that algorithmic systems align with the values of equity and justice, rather than being a reflection of inequalities that they are intended to resolve.[35] Laws are required to ensure that there is transparency within these tools and public trust is maintained. If companies keep the methods of how they train and apply data within algorithms hidden, it becomes difficult to build public trust in such systems.

As seen in the U.S. *Loomis* case, challenging the accuracy of algorithmic tools becomes extremely difficult when their training and deployment methods are hidden.[36] Algorithmic reparation would include the requirement of independent reviews to take place in order to showcase that these tools are transparent. Additionally, within the legal reform there should be the integration of historical and systemic considerations into the design and application of these data driven tools. This can include potentially embedding Indigenous-specific principles, such as *Gladue* principles, which focus on the need to account for systemic factors during sentencing decisions.[37] Currently, algorithmic tools and the constitutional law fail to address such nuanced frameworks, which results in treating marginalized groups as standardised data points.[38]

Bias auditing is also a critical component within algorithmic reparation, as it would be mandated by law to examine and address discriminatory patterns in algorithms.[39] These audits are an important implementation in order to foster transparency and public trust within the criminal justice system. Having bias auditing within the legal framework for algorithms will limit the feedback loops of bias inherent in predictive policing and risk assessment tools, which is reliant on historical biases like carding practices.[40] All in all, algorithmic reparation involves using legal, technical and social efforts to change the fairness and transparency levels in these tools that are influenced by historical biased data. Ultimately, through legislative reform, algorithmic reparation can ensure that data-driven tools are serving equitable outcomes rather than continuing systemic biases.

Conclusion

To conclude, algorithms and data-driven tools are far from being neutral tools, instead they continue to spread existing biases. Data-driven tools like risk assessments, sentencing algorithms, and predictive policing are designed to be fair and unbiased. However, this research has showcased that these algorithmic technologies are far from being neutral. Algorithmic reparation, through the integration of bias audits, historical considerations, and mandatory transparency, can help decrease the existing biases fixated in data-driven decision-making. Ideally, algorithmic reparation seeks to transform the criminal justice system into a more inclusive system, where technology serves to correct, rather than continue implementing existing inequalities.

[34] Davis, *supra* note 27 at 8
[35] *Ibid* at 4
[36] Christian, *supra* note 4 at 27
[37] *Gladue*, *supra* note 32 at para 69
[38] Huq, *supra* note 8 at 47
[39] Bagaric, *supra* note 1 at 142
[40] *Ibid* at 108

References

Jurisprudence

Ewert v. Canada, 2018 SCC 30, [2018] 2 S.C.R. 165

R. v. Gladue, [1999] 1 S.C.R. 688

R. v. Le, 2019 SCC 34, [2019] 2 S.C.R. 692

Books

Benjamin, Ruha. *Race After Technology: Abolitionist Tools for the New Jim Code* (2019). Cambridge, UK: Polity Press.

Articles

Bagaric, Mirko et al. "The Solution to the Pervasive Bias and Discrimination in the Criminal Justice System: Transparent and Fair Artificial Intelligence." (2022) 59:1 American Criminal Law Review 95

Christian, Gideon, "Legal Framework For The Use Of Artificial Intelligence (AI) Technology In The Canadian Criminal Justice System" (2024) 21:2 Canadian Journal of Law and Technology 109 – 135.

Crenshaw, Kimberle, "Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics" (1989) *University of Chicago Legal Forum* 139.

Davis, Jenny et al, "Algorithmic Reparation" (2021) 8 *Big Data & Society*

Huq, Aziz Z, "Racial Equity in Algorithmic Criminal Justice" (2019) 68:6 Duke Law Journal 1043.

Office of the Correctional Investigator. *Annual Report 2021-2022* (2021) Ottawa: Office of the Correctional Investigator.

Chapter 8

Reproductive Technology In The Age Of Artificial Intelligence: Bioethical And Legal Dilemmas In Fertility Treatments

Shizza Malik (she/her)

JD Candidate, Lincoln Alexander School of Law

Abstract

The integration of artificial intelligence ("AI") into fertility treatments has the potential to revolutionize reproductive medicine, particularly in processes like embryo selection and in vitro fertilization ("IVF") success prediction. However, these advancements raise complex bioethical and legal challenges that intersect with fertility law. This paper briefly comments on the ethical dilemmas posed by AI in fertility treatments, focusing on key issues such as informed consent, the commodification of embryos, algorithmic bias, the reinforcement of societal stereotypes and equity in access to care. It examines the tensions between medical innovation and patient autonomy, alongside the regulatory void surrounding AI-driven fertility treatments. Through a critical analysis of current legal frameworks, this paper proposes a regulatory approach to address these challenges, ensuring that AI tools in fertility care are implemented ethically while safeguarding patient rights. The proposed regulatory model would emphasize transparency, accountability and the equitable access to resources, with a particular focus on data privacy, algorithmic fairness and informed consent. Ultimately, this paper advocates for a legal response that balances innovation with ethical responsibility, ensuring that AI advancements in fertility law align with fundamental principles of justice, autonomy and equality.

Keywords: Artificial Intelligence, Embryo, Fertility, Health, Equity

Introduction

In the modern era of technological advancement, integrating artificial intelligence ("AI") into reproductive health has generated both fascination and intense debate, particularly in fertility treatments. AI technologies, including those used in in vitro fertilization ("IVF") for embryo selection, IVF success prediction and genetic trait selection, promise to revolutionize reproductive medicine by enhancing efficacy and improving outcomes for families facing infertility.[1] These advancements offer unprecedented opportunities, such as improved genetic fitness assessments and higher success rates.[2] However, the rapid progression of AI in reproductive medicine raises significant ethical and legal concerns. Issues such as informed consent, algorithmic bias and the commodification of embryos present risks of exploitation, discrimination and infringement on patient autonomy.[3] This essay contends that while the potential of AI in fertility treatments is transformative, strict regulation is essential. Focusing on informed consent, algorithmic bias and equitable access, such regulation will safeguard patient rights, ensure justice and promote ethical AI use in reproductive health.

Background

To understand the implications of AI in fertility treatments, key concepts must be defined, and its current applications examined. AI refers to machine learning processes where algorithms analyze data to make predictions or decisions.[4] In fertility medicine, AI optimizes outcomes and improves IVF success rates by selecting viable embryos. IVF involves fertilizing eggs outside the body and implanting embryos into the uterus.[5] Informed consent is crucial in these procedures, requiring patients to understand risks, benefits and alternatives.[6]

However, the complexity of AI can obscure its role in treatment. Algorithmic bias, where prejudiced training data leads to discriminatory outcomes, may reinforce stereotypes or disadvantage certain groups.[7] The commodification of embryos—treating them as commercial products—raises ethical questions about human dignity and exploitation.[8] AI tools, like embryo screening algorithms, show promise, with studies reporting up to a 30% improvement in success rate.[9] Yet, rapid adoption outpaces regulation,[10] highlighting the urgent need for clear, responsible guidelines in reproductive care.

Algorithmic Bias And Reinforcement of Stereotypes

Algorithmic bias in AI-driven fertility treatments is a pertinent ethical concern. AI systems are developed by humans and programmers who naturally hold inherent biases prevalent in society.

---

[1] David B Olawode et al, Artificial intelligence in in-vitro fertilization (IVF): A new era of precision and personalization in fertility treatments (2025) 54:1 Journal of Gynecology Obstetrics and Human Reproduction 1-11.
[2] Sahil Gupta, AI will revolutionize assisted reproductive technology (if we work together) (July 1, 2020) online: Fertility and Sterility <https://www.fertstert.org/news-do/ai-will-revolutionize-assisted-reproductive-technology-if-we-work-together>
[3] Julian J Koplin et al, Ethics of artificial intelligence in embryo assessment: mapping the terrain (2025) 40:2 Human Reproduction 179 at 185.
[4] Mikke F Kragh & Henrik Karstoft, "Embryo Selection with Artificial Intelligence: How to Evaluate and Compare Methods?" (2021) 38 J Assisted Reproduction & Genetics 1675 at 1676.
[5] *Ibid* at 1675.
[6] *Ibid.*
[7] Riikka Homanen, Neil McBride & Nicky Hudon, "Artificial Intelligence and Assisted Reproductive Technology: Applying a Reproductive Justice Lens" (2024) 31 European J Women's Studies 262.
[8] *Ibid*.
[9] M Salih et al, "Embryo Selection through Artificial Intelligence versus Embryologists: A Systematic Review" (2023) 3 Human Reproduction Open 2.
[10] S*upra* note 4.

These biases often translate into AI algorithms that support fertility treatments.[11] For instance, gender selection algorithms might be developed with a patriarchal bias towards selecting male embryos.[12] This interferes with normal reproduction and continues to circulate unconstructive gender stereotyping, reinforcing detrimental social inequalities that AI should actually be eliminating.[13] AI gender selection could exacerbate a patriarchal system even before individuals are birthed, leading to an unfavourable ratio of male-to-female births in cultures that favour boys, subordinate women and uphold gender roles, consequently fuelling social injustice.

Commodification of Embryos And Patient Autonomy

The integration of AI into fertility treatments introduces significant ethical concerns, particularly regarding the commodification and dehumanization of embryos. For instance, the increase in AI diagnostic capacity could lead to "designer babies", whereby scientists edit genetic information to create human who would exhibit specific physiological characteristics.[14] This practice risks reducing embryos to material objects and data points that can be created, selected or discarded based on subjective genetic preferences.[15] Accordingly, reports have suggested that there are at least 90,000 frozen embryos considered abandoned in the US, including some suggesting that the figures are in the millions which are then subsequently discarded.[16] Moreover, the disposal of surplus embryos, whether through destruction or donation for research, is an inherent part of the IVF process. This attempts to maximize the chances of success and minimize the need for repeated invasive procedures, as some embryos may stop growing, have chromosomal abnormalities or remain unused once patients have completed building their families. [17] With increased AI capacity, surplus and disposal of embryos may increase.

Such dehumanization contradicts the ethical view by some quarters that embryos, as potential human beings, deserve respect and protection.[18] Furthermore, it undermines the altruistic ideals of patient autonomy by transforming the deeply personal process of family creation into a commercialized fertility market. This commodification directly conflicts with the *Assisted Human Reproduction Act*, which prohibits the buying and selling of embryos in an effort to safeguard human dignity in reproductive medicine.[19]

---

[11] Lama H Nazer et al, "Bias in Artificial Intelligence Algorithms and Recommendations for Mitigation" (2023) 2 PLOS Digital Health e0000278.
[12] Natalia Norori et al, "Addressing Bias in Big Data and AI for Health Care: A Call for Open Science" (2021) 2:10 Patterns 1.
[13] S*upra* note 11.
[14] See for example Assisted Human Reproduction Act 2004, S C 2004. C 2, s 2(f),(g); "trade in the reproductive capabilities of women and men and the exploitation of children, women and men for commercial ends raise health and *ethical concerns that justify their prohibition*; and human individuality and diversity, and the *integrity of the human genome, must be preserved and protected*." [Emphasis added].
[15] See Chaira Longoni, Andrea Bonezzi and Carey K. Monrewedge, Resistance to Medical Artificial Intelligence (2019) 46 Journal of Consumer Research 629 at 630 where the authors argued against AI in medicine because it can lead to what they termed 'uniqueness neglect" for its failure to account for unique characteristics and circumstances.
[16] Mary Pflum, Nation's fertility clinics struggle with a growing number of abandoned embryos (August 12, 2019) online: NBC News <https://www.nbcnews.com/health/features/nation-s-fertility-clinics-struggle-growing-number-abandoned-embryos-n1040806>
[17] See Rachael Robertson, Why Discarding Embryos Is Inherent to the IVF Process (February 28, 2024) online: MedpageToday <https://www.medpagetoday.com/obgyn/infertility/108932>
[18] Michael J Sandel, Embryo Ethics — The Moral Logic of Stem-Cell Research (2004) 351:3 New England Journal of Medicine 207 – 209. The author presents varying views about the ethics of human embryos in stem-cell research.
[19] *Assisted Human Reproduction Act*, SC 2004, c 2, s 12; s. 2(f)(g) of the Act provides that; f) trade in the reproductive capabilities of women and men and the exploitation of children, women and men for commercial ends raise health and *ethical concerns that justify their prohibition*; and human individuality and diversity, and the *integrity of the human genome, must be preserved and protected*. [Emphasis added]. See also Julian J Koplin et al, Ethics of artificial intelligence in embryo assessment: mapping the terrain (2025) 40:2 Human Reproduction 179 at 180; the authors linked the "dehumanization" argument mainly to lack of patient consent rather than the human embryo.

Although patients are empowered to make reproductive choices, AI can introduce subtle, yet powerful pressures that erode autonomy and informed consent. Societal preferences for traits like blue eyes or specific physical attributes may influence AI-driven recommendations, pushing patients to select embryos conforming to these biases. This risks exploiting communities with historically underrepresented traits, such as green eyes, reinforcing harmful stereotypes and marginalization. Beyond appearance, the influence of AI extends to characteristics like intelligence, height or athleticism, perpetuating norms about "ideal" qualities and creating undue pressures to conform.

While AI is marketed to enhance patient autonomy by providing more options and information, it paradoxically undermines self-determination.[20] Patients may feel compelled to align their decisions with societal expectations embedded in AI systems rather than make independent choices. To protect patient rights and uphold ethics, regulations must prevent embryo exploitation, mitigate algorithmic bias and shield patients from societal pressures, ensuring responsible use of AI in fertility treatments.

Marxist Perspective On Able Bodies And Economic Growth

A Marxist critique of AI-assisted genetic enhancement reveals how these practices align with capitalist ideals, prioritizing economic productivity over human dignity. By valuing individuals based on traits associated with efficiency, such as physical strength or intellect, AI could be used to reinforce social and racial hierarchies.[21] This approach perpetuates a societal framework where individuals are seen as valuable only if they meet certain criteria for economic utility, such as their capacity for labour or contribution to productivity. This practice directly conflicts with the *Assisted Human Reproduction Act*, which seeks to protect human dignity,[22] but instead fosters social injustice and alienates those who cannot conform to these narrowly defined abilities.

The economic implications of AI-driven genetic selection also highlight the exclusionary potential of this technology. By privileging specific genetic traits, AI may perpetuate prejudice against embryos with forecasted disabilities or neurodivergent traits such as characteristics on the Autism Spectrum. Such practices reflect an ableist worldview, where only certain lives are deemed worth living, further marginalizing those who do not fit societal norms.[23] Disabled individuals may face heightened barriers to economic participation and social recognition, exacerbating the discrimination they already endure. This focus on genetic "perfection" risks erasing diversity and perpetuating prejudice. By privileging able-bodied traits, AI-assisted genetic enhancement promotes a narrow definition of worth that devalues minority expressions of humanity. This capitalist-driven pursuit of productivity reduces individuals to their economic potential, stripping away the richness of human diversity and reinforcing systemic inequality.

Equity And Access To AI-Driven Fertility Care

Studies show that IVF fertility treatments with genetic enhancements cost an average of USD 30,000,[24] a high cost that most middle-class families cannot afford. As these enhancements and

[20] Eduardo Hariton et al, "Applications of Artificial Intelligence in Ovarian Stimulation: A Tool for Improving Efficiency and Outcomes" (2023) 120:1 Fertility & Sterility 8 at 16.
[21] Miguel Beriain, "Human Dignity and Gene Editing" (2020) 19:10 EMBO Reports 1-4.
[22] *Assisted Human Reproduction Act*, *supra* note 14.
[23] Kristen Lyall et al, "Fertility Therapies, Infertility and Autism Spectrum Disorders in the Nurses' Health Study II" (2020) 26:4 Paediatric & Perinatal Epidemiology 361.
[24] Arian Khorshid et al, "Average Cost Of In Vitro Fertilization with Preimplantation Genetic Testing For Monogenic Disorders And Aneuploidy Per Unaffected Live Birth For Carrier Couples" (2021) 116:3 Fertility & Sterility e374.

treatments become more complex and AI is able to suggest or make higher level genetics enhancements, the cost would also increase exponentially. The high cost of AI-based fertility treatments poses significant barriers to equitable access, restricting these advancements to the wealthiest individuals. These cutting-edge technologies remain largely inaccessible for middle and low-income families, deepening existing healthcare inequalities.[25] This divide is particularly stark in developing countries, where such treatments are almost exclusively reserved for the upper class, leaving the majority without access to improved technological reproductive care.[26] This inequity creates a dual system of fertility healthcare—one offering state-of-the-art treatments for those who can afford them, and another where families are left struggling due to a lack of resources.

The global disparity in access to AI-enhanced reproductive medicine further widens the gap between developed and developing nations, solidifying structural inequities.[27] Marginalized groups, many of whom are already underrepresented in healthcare, face disproportionate exclusion from these advancements. Without efforts to expand access, AI-driven fertility care risks worsening the disparities in reproductive healthcare and infringing on reproductive freedoms. Addressing this issue requires policies aimed at ensuring these innovations are accessible to all, regardless of income or geography, to prevent further entrenchment of global health inequities and to protect the principle of equal access to reproductive care.

Counterarguments and Refutations

a. AI as a Tool for Medical Innovation

AI has the potential to transform fertility treatments by enhancing the success rates of IVF procedures. Through big data analysis, AI can predict the likelihood of viable embryos, reducing the financial and emotional toll of repeated IVF cycles.[28] However, the benefits of AI cannot be divorced from its ethical and societal implications. The risk of biases embedded in AI systems persists, leading to the commodification of embryos and unequal access to treatments. These issues disproportionately affect marginalized communities, compounding existing inequities. To ensure AI-driven innovations align with ethical principles, robust regulations are essential to protect against exploitation and injustice.

b. Parental Autonomy in Genetic Choices

While individuals have the right to make reproductive decisions, unchecked autonomy in genetic selection poses significant societal risks. Choosing traits such as intellect or immunity to diseases could normalize discrimination, reinforce societal preferences for certain qualities and commodify human life.[29] The *Assisted Human Reproduction Act* explicitly deems genetic selection for non-medical reasons unethical, emphasizing that human life should not be reduced to a product of economic or societal preferences.[30] Moreover, informed consent remains a critical legal safeguard. The *Canadian Medical Association's Code of Ethics* mandates that physicians provide comprehensive information about medical procedures, including AI-based fertility treatments[31]. This principle is

[25] Supra note 7 at 262.
[26] Ahmed Shahin, The problem of IVF cost in developing countries: has natural cycle IVF a place? (2007) 15:1 Reproductive BioMedicine Online 51 at 54 – 55.
[27] *Ibid* at 262.
[28] Darren Chow et al, "Does Artificial Intelligence Have a Role in the IVF Clinic?" (2021) 2:3 Reproduction & Fertility C29.
[29] Giulia Cavaliere, "The Problem with Reproductive Freedom: Procreation Beyond Procreators' Interests" (2020) 23:1 Medicine, Health Care & Philosophy 131.
[30] *Assisted Human Reproduction Act*, *supra* note 14.
[31] Canadian Medical Association, *Code of Ethics and Professionalism* (Ottawa: The Association, 2018), s. 35

reinforced in *Malette v. Shulman (1990)*, where the Supreme Court affirmed that patients must have the autonomy to make healthcare decisions free from coercion[32]. This legal precedent underscores the need for true transparency and ethical standards to guide the role of AI in reproductive medicine.

Proposed Regulatory Framework

a. Transparency and Accountability

Transparency is essential for regulating AI-driven fertility treatments to build patient trust and engagement. Developers and fertility clinics must disclose how data is collected, analyzed and used to predict embryo viability or IVF success rates.[33] Patients should have clear access to this information to make fully informed decisions about their care. Limiting patients' ability to opt out of data sharing must be accompanied by safeguards to protect their privacy and reduce risks. This approach upholds the principle of informed consent and complies with the *Personal Information Protection and Electronic Documents Act,* which mandates accountability and transparency in handling personal data.[34]

b. Algorithmic Fairness

To prevent perpetuating bias, AI systems must be trained on diverse datasets that reflect the full range of human variation. Such diversity ensures that algorithms do not favour specific populations or reinforce societal inequalities[35]. Regular audits by independent and unbiased bodies should assess and rectify any unfair outcomes AI models produce. These measures will ensure that AI technologies contribute to fairness and justice in fertility treatments, avoiding harm or prejudice against marginalized groups.

c. Informed Consent

Fertility clinics must provide patients with clear and comprehensive information about how AI technologies influence their treatments. This includes outlining the purpose, benefits, limitations and potential risks of AI systems. Such transparency is mandated by the *Canadian Medical Association's Code of Ethics*, which emphasizes patient autonomy and informed decision-making[36]. Clinics should also offer resources to educate patients about these technologies in accessible terms, ensuring they are equipped to provide meaningful consent without pressure or coercion.

d. Equitable Access to Care

Financial barriers to AI-driven fertility treatments must be addressed to promote equitable access. Subsidies, funding programs or expanded insurance coverage are critical to ensure that low-income families can access these technologies.[37] This aligns with the *Canada Health Act*, which guarantees equal access to healthcare for all Canadians. By reducing the costs of AI-based reproductive care, such measures would bridge existing gaps and uphold reproductive justice, ensuring that no group is excluded from the benefits of medical innovation. This framework ensures that AI-driven

[32] Norman Siebrasse, "Malette v. Shulman: The Requirement of Consent in Medical Emergencies" (1989) 34:4 McGill LJ 1080
[33] S*upra* note 29.
[34] *Personal Information Protection and Electronic Documents Act,* SC 2000, c 5.
[35] S*upra* note 11.
[36] S*upra* note 29.
[37] S*upra* note 7.

fertility treatments operate ethically, equitably and transparently, protecting patient rights and promoting justice in reproductive healthcare.

Conclusion

AI-driven fertility treatments raise profound ethical and legal challenges, including concerns about informed consent, algorithmic bias and equitable access. Without regulation, these technologies risk exacerbating societal inequalities, commodifying human life and undermining autonomy. A robust regulatory framework will ensure transparency, fairness and dignity, safeguarding patient rights while fostering responsible innovation in reproductive healthcare.

References

Legislation

Assisted Human Reproduction Act, SC 2004, c 2.

Canadian Human Rights Act, RSC 1985, c H-6.

Canadian Medical Association Code of Ethics and Professionalism (2018).

Personal Information Protection and Electronic Documents Act, SC 2000, c 5

Secondary Materials: Articles

Beriain, Miguel, "Human Dignity and Gene Editing" (2020) 19:10 EMBO Reports 1 – 4.

Cavaliere, Giulia, "The Problem with Reproductive Freedom: Procreation Beyond Procreators' Interests" (2020) 23:1 Medicine, Health Care & Philosophy 131.

Chow, Darren et al, "Does Artificial Intelligence Have a Role in the IVF Clinic?" (2021) 2:3 Reproduction & Fertility C29.

Hariton, Eduardo et al, "Applications of Artificial Intelligence in Ovarian Stimulation: A Tool for Improving Efficiency and Outcomes" (2023) 120 Fertility & Sterility 8.

Homanen, Riikka, Neil McBride & Nicky Hudon, "Artificial Intelligence and Assisted Reproductive Technology: Applying a Reproductive Justice Lens" (2024) 31 European J Women's Studies 262.

Khorshid, Arian et al, "Average Cost of In Vitro Fertilization with Preimplantation Genetic Testing For Monogenic Disorders And Aneuploidy Per Unaffected Live Birth For Carrier Couples" (2021) 116:3 Fertility & Sterility .

Koplin Julian J, et al, Ethics of artificial intelligence in embryo assessment: mapping the terrain (2025) 40:2 Human Reproduction 179 at 185.

Kragh, Mikke F & Henrik Karstoft, "Embryo Selection with Artificial Intelligence: How to Evaluate and Compare Methods?" (2021) 38 J Assisted Reproduction & Genetics 1675.

Longoni, Chaira, Andrea Bonezzi and Carey K. Monrewedge, Resistance to Medical Artificial Intelligence (2019) 46 Journal of Consumer Research 629 at 630.

Lyall, Kristen et al, "Fertility Therapies, Infertility and Autism Spectrum Disorders in the Nurses' Health Study II" (2020) 26 Paediatric & Perinatal Epidemiology 361.

Nazer, Lama H et al, "Bias in Artificial Intelligence Algorithms and Recommendations for Mitigation" (2023) 2 PLOS Digital Health e0000278.

Norori, Natalia et al, "Addressing Bias in Big Data and AI for Health Care: A Call for Open Science" (2021) 2:10 Patterns 100347.

Olawode, David B, et al, Artificial intelligence in in-vitro fertilization (IVF): A new era of precision and personalization in fertility treatments (2025) 54:1 Journal of Gynaecology Obstetrics and Human Reproduction 1-11.

Salih, M et al, "Embryo Selection through Artificial Intelligence versus Embryologists: A Systematic Review" (2023) Human Reproduction Open [insert first page].

Sandel, Michael J, Embryo Ethics — The Moral Logic of Stem-Cell Research (2004) 351:3 New England Journal of Medicine 207 – 209. The author presents varying views about the ethics of human embryos in stem-cell research.

Shahin, Ahmed, The problem of IVF cost in developing countries: has natural cycle IVF a place? (2007) 15:1 Reproductive BioMedicine Online 51 at 54 – 55.

Siebrasse, Norman, "Malette v. Shulman: The Requirement of Consent in Medical Emergencies" (1989) 34:4 McGill L J 1080.

Online: Websites

Gupta, Sahil, AI will revolutionize assisted reproductive technology (if we work together) (July 1, 2020) online: Fertility and Sterility https://www.fertstert.org/news-do/ai-will-revolutionize-assisted-reproductive-technology-if-we-work-together

Pflum, Mary, Nation's fertility clinics struggle with a growing number of abandoned embryos (August 12, 2019) online: NBC News https://www.nbcnews.com/health/features/nation-s-fertility-clinics-struggle-growing-number-abandoned-embryos-n1040806

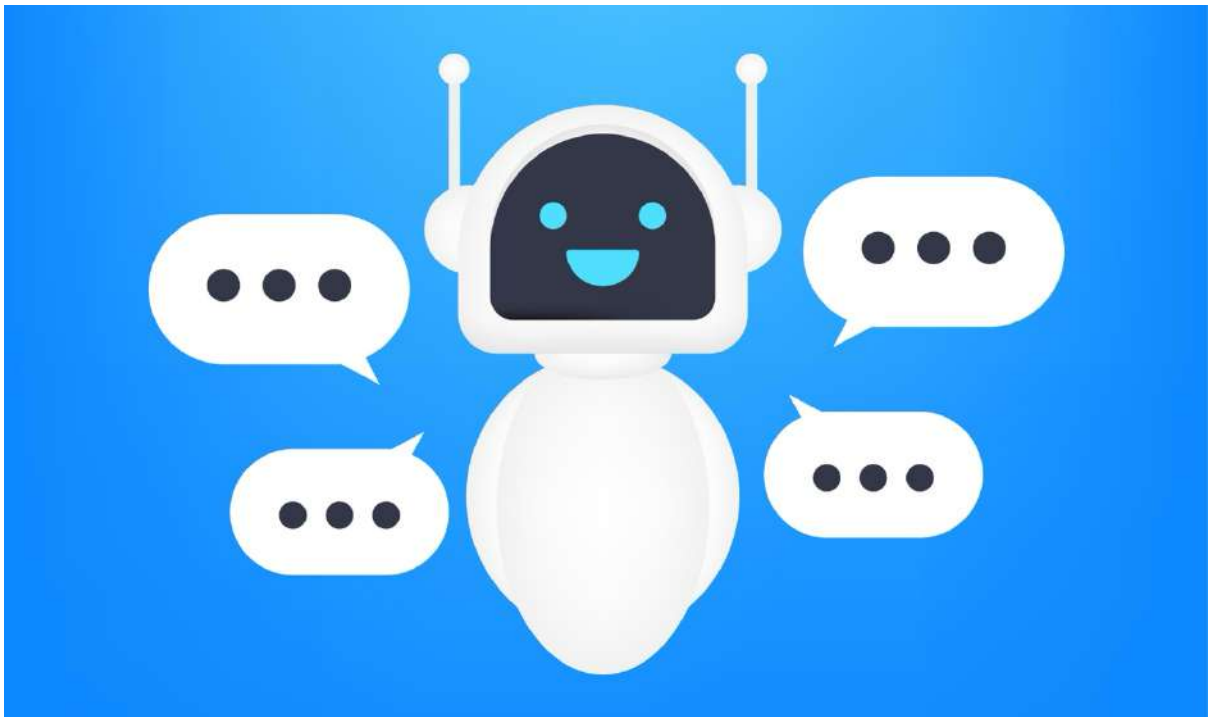
Robertson, Rachael, Why Discarding Embryos Is Inherent to the IVF Process (February 28, 2024) online: MedpageToday <https://www.medpagetoday.com/obgyn/infertility/108932>

Chapter 9

Sweet Dream or a Beautiful Nightmare? Artificial Intelligence Chatbots for Self-Represented Family Law Litigants

Julia Marr (She/Her)

JD Candidate, Lincoln Alexander School of Law



Abstract

In this short paper, I explore the possibilities of AI Chatbots to improve access to justice for self-presented litigants. Analysis weighs technological determinism with the critical social realities of family law proceedings to make meaningful recommendations for government agencies and the judicial system.

Keywords: AI chatbots, family law, feminist law, disability law, accessibility, access to justice.

Introduction

Family law is a major site of stress in the Canadian legal system.[1] Self-represented family law litigants[2] face unique challenges. Many do not have the education or literacy skills to benefit from the voluminous legal information found online, and some have visual impairments or other disabilities that make them inaccessible.[3] With Artificial intelligence ("AI"), there is a potential to provide clear and concise answers to self-represented litigants' questions via chatbots. In this paper, I first discuss the problem of access to justice in family law. Second, I discuss how AI chatbots can promote access to justice by providing accessible and individualized legal information. I analyse AI from the perspective of racial justice, disability justice, and intersectional feminism I conclude by reflecting on the risks of using AI in in family law.

My Lived Experiences in Family Law

At 8 years old, I was apprehended by the Barrie Police Service ("BPS") as a part of a child protection investigation. My removal by BPS was intrusive and frightening. I then lived in a foster home and was later put into a kinship placement. Meanwhile, my mother sought the advice of a lawyer, who allowed her and our family to move on from the situation. My lived experiences influence how I view the family law system, with hopeful scepticism. In this paper, I propose ways to improve access to justice to benefit those interacting with the family law system.

The Access to Justice Problem in Family Law

Family law problems are seen in significant numbers. From 2022 to 2023, there were 252,516 active family law cases across Canada.[4] High-conflict family law disputes take 27.7 months on average to litigate[5]. Within three years, 1,216,497 Canadians (5.1 percent of the adult population) reported experiencing a family law problem that was "serious ... and not easy to fix."[6] A full third of all non-criminal cases heard in Canadian courts are family law cases.[7]

Over half of the family cases in Canada's courts now have one or both parties without a lawyer.[8] Financial reasons and ineligibility for legal aid are the most significant explanatory factors for the lack of legal representation.[9] For some litigants, the decision to self-represent reflects a confidence in their knowledge and ability to navigate the system, a distrust of lawyers or a desire to deal directly with their former partner.[10]

Whatever the reason parties decide to represent themselves, they need accessible and helpful legal information. Legal actors' legislative duty to provide complete, accurate and up-to-date legal

[1] Federal Judicial Center, "Federal Judicial Caseloads, 1789-2016: Trial Court Caseloads since 1870" (no date), online: Federal Judicial Center <www.fjc.gov/history/exhibits/graphs-and-maps/trial-court-caseloads-1870>.
[2] The term "self-represented" describes persons who appear before the court without representation from a lawyer.
[3] Rachel Birnbaum, Nicholas Bala and Lorne D Bertrand, "The Rise of Self-Representation in Canada's Family Courts: The Complex Picture Revealed in Surveys of Judges, Lawyers and Litigants", (2013) 91:1 Canadian Bar Review 68.
[4] Statistics Canada, "Active family cases by issue(s) identified over length of case and number of fiscal years since case initiation, Canada and selected provinces and territories" (2024) Civil Court Survey Table 35-10-0113-01.
[5] Joanne J Paetsch, Lorne D Bertrand and John-Paul E Boyd, "An Evaluation of the Cost of Family Law Disputes: Measuring the Cost Implication of Various Dispute Resolution Methods", (2018) Canadian Forum on Civil Justice 16.
[6] Trevor CW Farrow et al, "Everyday Legal Problems and the Cost of Justice in Canada: Overview Report" (2016) Canadian Forum on Civil Justice 7.
[7] Mary Bess Kelly, "Divorce Cases in Civil Court", (2010/2011) Canadian Centre for Justice Statistics 15.
[8] Birnbaum, *supra* note 3 at 71.
[9] Birnbaum, *supra* note 3 at 76.
[10] Birnbaum, *supra* note 3 at 76.

information to litigants.[11] Lawyers must be courteous, civil, and act in good faith with self-represented parties.[12]

Online legal information is so voluminous that it is inaccessible.[13] In a 2013 survey, only 12% of family law self-represented litigants found information on the Ministry of Attorney General ("MAG") to be very helpful, and 47% found it to be somewhat beneficial.[14]

Technological Transformation

Established by then-Chief Justice Beverley McLachlin in 2007, the Action Committee on Access to Justice in Civil and Family Matters ("the Action Committee") works towards improved access to justice for people in Canada. In 2013, the Action Committee recommended that all justice system stakeholders support the exploration of the potential for the Internet and information technology to make family justice more affordable and accessible.[15] This recommendation was made over 10 years ago and has yet to be implemented.

The COVID-19 pandemic exemplifies how Canadian courts can accelerate the digital transformation. Within a fortnight, there was a technological upheaval as the justice system moved from a world in which almost all court hearings were held in person to one in which nearly none were.[16] While many judges and lawyers may have initially had visceral negative reactions to the prospect of virtual hearings, they quickly adapted.[17] The pandemic caused significant changes to the family law system; litigants can serve others and be served by email[18], and some Court appearances can be held virtually[19]. The family law system must maintain the same level of technological innovation seen in the pandemic to increase access to justice.

The Possibilities of AI Chatbots

AI refers to systems that display intelligent behaviour by analyzing their environment and taking actions – with some degree of autonomy – to achieve specific goals.[20] AI-powered chatbots simulate human conversation through text or voice using algorithms to process and understand natural language.[21] Chatbots provide instant, personalized responses to users' inquiries.[22]

---

[11] *Divorce Act*, RSC 1985, s. 7.4, *Rules of Professional Conduct* Rule 3.1-1.
[12] *Rules of Professional Conduct*, r. 7.2-9.
[13] Julie Macfarlane, "National (Canada) Self-Represented Litigants Final Report: Identifying and Meeting Needs of Self-Represented Litigants" (2013) at 31, 91-92.
[14] Birnbaum, *supra* note 3 at 86.
[15] Canada, The Action Committee on Access to Justice in Civil and Family Matters, *Meaningful Change for Family Justice: Beyond Wise Word,* (Ottawa 2013), online: <https://www.cfcj-fcjc.org/sites/default/files/docs/2013/Report%20of%20the%20Family%20Law%20WG%20Meaningful%20Change%20April%202013.pdf>
[16] Richard Susskind, *The Future of Courts* (United States of America: Harvard Law School Center on the Legal Profession, 2020), online: <https://clp.law.harvard.edu/knowledge-hub/magazine/issues/remote-courts/the-future-of-courts/>.
[17] Samuel Dahan and David Liang , "The Case for AI-Powered Legal Aid" (2021) Queen's Law Journal 418.
[18] Superior Court of Justice, "Notice to Profession – Toronto Expansion Protocol for Court Hearings During COVID-19 Pandemic" (Ontario: Superior Court of Justice, 2022) online: <https://www.ontariocourts.ca/scj/notices-and-orders-covid-19/notice-to/>.
[19] Action Committee on Modernizing Court Operations. *Virtual and Hybrid Hearings in Family Matters: Promoting the Best Interests of the Child* (Ontario: Action Committee on Modernizing Court Operations, 2024) online: https://www.fja.gc.ca/COVID-19/pdf/Childrens-Interests-in-Virtual-Hearings.pdf
[20] The European Commission's Communication, *A definition of AI: Main capabilities and scientific disciplines*, (Brussels, European Commission, 2018) online: <https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_december_1.pdf>.
[21] Nze, Stella Udoka. "AI-Powered Chatbots" Global Journal of Human Resource Management, vol. 12, no. 6, at 34–45.
[22] *Ibid*, at 34–45.

I propose that the federal and provincial governments create an AI chatbot to answer self-represented litigants' questions with legal information. Collaboration between the different levels of government will ensure that unmarried and married family law litigants can rely on the chatbot equally.[23] Government entities typically remain free from the impacts of privatization and monetization, allowing the chatbot to remain a matter of public interest.

In addition to receiving legal information, users would be directed to the appropriate forms to help resolve their family law issues.[24] This would streamline the voluminous and often unhelpful family law information found online.[25] In addition, ensure self-represented litigants receive correct legal information.

Unique to this chatbot would be the focus on mediation, anti-oppressive lawyering, and trauma-informed lawyering. These perspectives have proven to be effective for self-represented litigants in seeking access to justice.[26] Lawyers have a legislative duty to encourage alternative dispute resolution processes, such as negotiation, mediation or collaborative law when it is appropriate to do so.[27] A trauma-informed family dispute resolution process recognizes the risk of harm, adjusts the process in response to signs of trauma, and has much to offer all family clients.[28]

Associated Risks

In AI ideology, there is often relentless technological optimism, the belief that technological progress is an autonomous force and can save us all, and the tendency to delegate key decisions to opaque algorithms.[29] To avoid such fallacies, I will explore the risks of the proposed chatbot: privacy, accessibility, and bias.

a. Privacy

Privacy is a significant concern. Family law contains personal information about identifiable individuals. For the chatbot to protect users' privacy, the government must uphold their legislative duties in the *Privacy Act*. This includes taking all reasonable steps to ensure that personal information used for an administrative purpose by the institution is as accurate, up-to-date and complete as possible.[30] In addition, properly disposing of personal information per the regulation.[31]

Unique to the chatbots is the concern of linking together information in novel ways that reveal sensitive information about other individuals who previously made entries.[32] Therefore, the government must closely monitor the chatbot. One concern may be that a chatbot could be used to further litigation harassment by seeking information about an opposing party who previously

[23] In Ontario, married couples are subject to the federal *Divorce Act,* RSC 1985, s. 3 and unmarried couples are subject to Ontario's *Family Law Act*, RSO 1990, s. 1(1).
[24] Ontario Court Forms is a website providing PDFs and word documents for various areas of law including the *Family Law Rules* Forms, *Children's Law Reform Ac*t forms, and Office of the Children's Lawyer forms.
[25] Birnbaum, *supra* note 3 at 89.
[26] Birnbaum, *supra* note 3 at 97: All respondents were also asked if they used the Ministry of Attorney General (MAG) website. Less than half (37%) reported that they had used the website; of those who had, 21% reported that it was very helpful and another 47% reported that it was somewhat helpful.
[27] *Children's Law Reform Act*, SO 1990, s. 33.1(3).
[28] Rhona Buchan, "Do No Harm: The Case for a Trauma-Informed Family Law Practice & the Use of Family Dispute Resolution", (2022) Journal of the Canadian Collaborative for Engagement & Conflict Management 137.
[29] Simon Lindgren, "Introducing Critical Studies of Artificial Intelligence" in ed, Handbook of Critical Studies of Artificial Intelligence, eds, (Cheltenham, UK: Edward Elgar Publishing, 2023) 3.
[30] *Privacy Act*, RSC 1985, c P-21, s. 6(2).
[31] *Ibid*, s. 6(3).
[32] Feder Cooper et al, "Report of the 1st Workshop on Generative AI and Law" (2023) Cornell University at 11.

used the chatbot. Minimizing litigation harassment in instances of intimate partner violence ("IPV") has been a significant focus of family law literature in recent years.[33]

b. Accessibility

For an AI chatbot to be effective, it must be accessible. Just 26% of Canadians believe the family law system is accessible.[34] Differences among people in terms of literacy, language skills, and internet access will mean that not everyone is likely to benefit from the introduction of technology.[35] The chatbot must be co-designed by disabled people, ensuring they are vital stakeholders. In addition, the chatbot should be created per accessibility regulations and best practices. For example, under the *Accessibility for Ontarians with Disabilities Act*, the Ontario government has a legislative duty to conform with the World Wide Web Consortium Web Content Accessibility Guidelines (WCAG) 2.0, at Level AA.[36]

c. Bias

Bias is also a concern. Bias emerges as a result of specific elements and decisions in the process of designing AI systems.[37] Like judges presiding over a case, technology can never be neutral, in the sense of purely objective, but they can and must strive for impartiality.[38] Therefore, those involved with designing the chatbot must refrain from bringing their own biases as it may negatively affect the efficacy of the chatbot. In addition, designers must scrutinize the materials and legal information the chatbot provides users. This will require careful consideration as family law jurisprudence and discourse is shifting away from incorrect notions of gender[39], race[40], class[41], and indigeneity[42]. The chatbot must include legislative reforms and contemporary case law.

A concern for chatbots is abuse by users. In South Korea, users manipulated a chatbot named "Lee Luda" to respond with overly sexualized, racist, homophobic, or misogynistic comments in response to specific prompts, to the extent that the chatbot made it to news headlines and the start-up ended up shutting down its service within days of the launch.[43] To avoid similar negative outcomes, the chatbot must be initially tested and continuously monitored to prevent abusive outcomes, as it will lower the efficacy of the chatbot.

d. Transparency

Lastly, the AI chatbot must be transparent. Users must know where the information comes from, which the chatbot can accomplish by providing citations and direct links to sources. To ensure

---

[33] Haya Sakakini, "Psychological Abuse Claims in Family Law Courts in BC: Legal Applications and Gaps" (2021) Canadian Journal of Family Law 34;1; Deanne M Sowter, "Intimate Partner Violence and Ethical Lawyering - Not Just Special Rules for Family Law" (2024) Canadian Bar Review 130.
[34] Ting Li, "Perceptions of and confidence in the Canadian family justice system: Key findings from the 2022 National Justice Survey" (2022) Research and Statistics Division Department of Justice Canada 5.
[35] Jane Bailey, Jacquelyn Burkell and Graham J Reynolds, "Access to Justice for all: Towards an "Expansive Vision" of Justice and Technology", (2013) Windsor Yearbook on Access to Justice 206.
[36] *Integrated Accessibility Standards*, O Reg 191/11, s. 14(1).
[37] Denis Newman-Griffis et al, "Definition drives design: Disability models and mechanisms of bias in AI technologies" (2023) First Monday 11.
[38] *Committee for Justice and Liberty et al. v. National Energy Board et al.,* 1976 CanLII 2 (SCC), [1978] 1 SCR 369 at page 385.
[39] United Nations, "Convention on the Elimination of All Forms of Discrimination Against Women" (2017) Committee on the Elimination of Discrimination Against Women, online: <https://documents-dds-ny.un.org/doc/UNDOC/GEN/N17/231/54/PDF/N1723154.pdf?OpenElement>.
[40] *Van de Perre v Edwards*, 2001 SCC 60 at para 41.
[41] Mary Jane Mossman et al, *Families and the Law: Cases and Commentary* (Canada: Captus Press, 2019) at 352.
[42] *An Act respecting First Nations, Inuit and Métis children, Youth and Families*, SC 2019, s. 8.
[43] Dongwoo Kim, "Chatbot Gone Awry Starts Conversations About AI Ethics in South Korea", *The Diplomat* (January 16, 2021), online: <thediplomat.com/2021/01/chatbot-gone-awry-starts-conversations-about-ai-ethics-in-south-korea>.

greater transparency, the Government could follow the 2019 California law requiring chatbots to disclose that they are not human.[44]

Conclusion

To conclude, the family law system must undergo a technological transformation to improve access to justice. In Canada, over 20% of the population take no meaningful action for their legal problems, and over 65% think that nothing can be done, are uncertain about their rights, do not know what to do, believe it will take too much time, cost too much money or are simply afraid.[45] This cannot continue. Governments must leverage technology to improve the legal system. The government must ensure regulation is sufficiently nimble and appropriately informed about technology to be effective.[46] If not, governments risk further ostracizing members of the public and lowering the credibility of the legal system.

[44] Renee DiResta, "A New Law Makes Bots Identify Themselves—That's the Problem" *Wired* (July 24 2019), online; Wired.com <https://www.wired.com/story/law-makes-bots-identify-themselves/>.
[45] Ab Currie, "The Legal Problems of Everyday Life: The Nature, Extent and Consequences of Justiciable Problems Experienced by Canadians" (2007) Department of Justice Canada 55.
[46] Jena McGill and Amy Salyzyn, "Judging by the Numbers: Judicial Analytics, the Justice System and its Stakeholders", (2021) Dalhousie Law Journal 279.

References

Jurisprudence

Committee for Justice and Liberty et al. v. National Energy Board et al., 1976 CanLII 2 (SCC), [1978] 1 SCR 369 at page 385.

Van de Perre v Edwards, 2001 SCC 60 at para 41.

Legislation

An Act respecting First Nations, Inuit and Métis children, Youth and Families, SC 2019, s. 8.

Children's Law Reform Act, SO 1990, s. 33.1(3).

Divorce Act, RSC 1985, s. 7.4, *Rules of Professional Conduct* Rule 3.1-1.

Integrated Accessibility Standards, O Reg 191/11, s. 14(1).

*Rules of Professional Conduc*t, r. 7.2-9.

Books

Lindgren, Simon ed, "Introducing Critical Studies of Artificial Intelligence" in Simon Lindgren ed *Handbook of Critical Studies of Artificial Intelligence* (Cheltenham, UK: Edward Elgar Publishing, 2023) 4.

Articles

Bailey, Jane, Jacquelyn Burkell and Graham J Reynolds, "Access to Justice for all: Towards an "Expansive Vision" of Justice and Technology", (2013) Windsor Yearbook on Access to Justice 206.

Birnbaum, Rachel, Nicholas Bala and Lorne D Bertrand, "The Rise of Self-Representation in Canada's Family Courts: The Complex Picture Revealed in Surveys of Judges, Lawyers and Litigants", (2013) 91-1 Canadian Bar Review 68.

Dahan, Samuel and David Liang , "The Case for AI-Powered Legal Aid" (2021) Queen's Law Journal 418.

Farrow, Trevor CW et al, "Everyday Legal Problems and the Cost of Justice in Canada: Overview Report" (2016) Canadian Forum on Civil Justice 7.

John Macfarlane, "National (Canada) Self-Represented Litigants Final Report: Identifying and Meeting Needs of Self-Represented Litigants" (2013) at 31, 91-92.

Kelly, Mary Bess, "Divorce Cases in Civil Court", (2010/2011) Canadian Centre for Justice Statistics 15.

McGill, Jena and Amy Salyzyn, "Judging by the Numbers: Judicial Analytics, the Justice System and its Stakeholders", (2021) Dalhousie Law Journal 279.

Mossman, Mary Jane et al, *Families and the Law: Cases and Commentary* (Canada: Captus Press, 2019) at 352.

Nze, Stella Udoka, "AI-Powered Chatbots" Global Journal of Human Resource Management, vol. 12, no. 6, at 34–45.

Paetsch, Joanne J, Lorne D Bertrand and John-Paul E Boyd, "An Evaluation of the Cost of Family Law Disputes: Measuring the Cost Implication of Various Dispute Resolution Methods", (2018) Canadian Forum on Civil Justice 16.

Sakakini, Hay, "Psychological Abuse Claims in Family Law Courts in BC: Legal Applications and Gaps" (2021) Canadian Journal of Family Law 34;1; Deanne M Sowter, "Intimate Partner Violence and Ethical Lawyering - Not Just Special Rules for Family Law" (2024) Canadian Bar Review 130.

International Documents

The European Commission's Communication, *A definition of AI: Main capabilities and scientific disciplines*, (Brussels, European Commission, 2018) online: <https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_december_1.pdf>.

United Nations, "Convention on the Elimination of All Forms of Discrimination Against Women" (2017) Committee on the Elimination of Discrimination Against Women, online: <https://documents-dds-ny.un.org/doc/UNDOC/GEN/N17/231/54/PDF/N1723154.pdf?OpenElement>.

Government Documents

Action Committee on Access to Justice in Civil and Family Matters, "Meaningful Change for Family Justice: Beyond Wise Words: Final Report of the Family Justice Working Group" (2013) Canadian Forum at 9 online: <https://www.cfcj-fcjc.org/sites/default/files/docs/2013/Report%20of%20the%20Family%20Law%20WG%20Meaningful%20Change%20April%202013.pdf>.

Action Committee on Modernizing Court Operations. *Virtual and Hybrid Hearings in Family Matters: Promoting the Best Interests of the Child* (2024) at 7 online: <https://www.fja.gc.ca/COVID-19/pdf/Childrens-Interests-in-Virtual-Hearings.pdf>.

Federal Judicial Center, "Federal Judicial Caseloads, 1789-2016: Trial Court Caseloads since 1870" (no date), online: Federal Judicial Center <www.fjc.gov/history/exhibits/graphs-and-maps/trial-court-caseloads-1870>.

Li, Ting, "Perceptions of and confidence in the Canadian family justice system: Key findings from the 2022 National Justice Survey" (2022) Research and Statistics Division Department of Justice Canada at 5 online: <https://www.justice.gc.ca/eng/rp-pr/jr/cfjs2022-sjfc2022/index.html>

Statistics Canada, "Active family cases by issue(s) identified over length of case and number of fiscal years since case initiation, Canada and selected provinces and territories" (2024) at Civil Court Survey Table 35-10-0113-01 online: <https://open.canada.ca/data/en/dataset/7ca04949-e6ef-4864-b133-a7e2149261c9/resource/f380bbe6-d51c-4156-9ff7-2b4150a5bb4e>.

Superior Court of Justice, "Notice to Profession – Toronto Expansion Protocol for Court Hearings During COVID-19 Pandemic" (Ontario: Superior Court of Justice, 2022) online: <https://www.ontariocourts.ca/scj/notices-and-orders-covid-19/notice-to/>.

The Action Committee, "About the Action Committee" (no date) online: <https://www.justicedevelopmentgoals.ca/about>.

Secondary Sources: Others

Currie, Ab, "The Legal Problems of Everyday Life: The Nature, Extent and Consequences of Justiciable Problems Experienced by Canadians" (2007) Department of Justice Canada 55.
Cooper et al, Feder, "Report of the 1st Workshop on Generative AI and Law" (2023) Cornell University at 11.

DiResta, Renee, "A New Law Makes Bots Identify Themselves—That's the Problem" *Wired* (July 24, 2019), online; Wired.com <https://www.wired.com/story/law-makes-bots-identify-themselves/>.

Dongwoo Kim, "Chatbot Gone Awry Starts Conversations About AI Ethics in South Korea", *The Diplomat* (January 16 2021), online: <thediplomat.com/2021/01/chatbot-gone-awry-starts-conversations-about-ai-ethics-in-south-korea>.

Newman-Griffis, Denis et al, "Definition drives design: Disability models and mechanisms of bias in AI technologies" (2023) First Monday 11.

Susskind, Richard, *The Future of Courts* (United States of America: Harvard Law School Center on the Legal Profession, 2020), online: <https://clp.law.harvard.edu/knowledge-hub/magazine/issues/remote-courts/the-future-of-court>.

Chapter 10

The Dangers of Dehumanizing the Loop: Applying a Critical Feminist Lens to Automated Weapons Systems

Grace McColl (she/her)

JD Candidate, Lincoln Alexander School of Law

Abstract

Military applications of artificial intelligence (AI), such as autonomous weapons systems (AWS) are increasingly being tested and deployed by states globally. While international debate persists about the benefits and risks of AWS and there is no binding international law on their use, there is consensus that AWS will increase both the frequency and intensity of armed conflict. This work applies a critical feminist lens (CFL) to the development and use of AWS in armed conflict to show how AWS may disproportionately harm those who do not fit Audre Lorde's "mythical norm". In addition to applying technofeminism to show how gender gaps in both the AI and defense sectors perpetuate and exacerbate existing inequities and biases, this paper leverages elements of feminist care ethics and relational theory, such as dehumanization, to illustrate the dangers of removing the 'human-in-the-loop' from AWS. Applying a CFL also expands notions of what constitutes 'harm' beyond the traditional physical harm standards used in military decision making and reveals the indirect consequences of AWS on the Global South, including environmental harm and exploitation of racialized women. The paper concludes with policy recommendations for addressing and regulating the use of AWS at both the international and national levels and increasing representation of intersectional individuals in the AI and defense sectors.

Keywords: armed conflict; artificial intelligence; automated weapons systems; critical feminist theory; international law

Introduction

Technology and militarism, both of which perpetuate patriarchal values, collude to position automated weapons systems (AWS) as tools that will exacerbate harms caused by armed conflict.[1] Applying a critical feminist lens (CFL) to the development and use of AWS shows how AWS may disproportionately harm those who do not fit Audre Lorde's mythical norm and expands the notion of what constitutes harm beyond physical harm. After contextualizing AWS and feminist approaches to artificial intelligence (AI) and militarism, this work argues that AWS would decrease compliance with international humanitarian law (IHL) and further dehumanize individuals already targeted by Western militarism. On a final note, the paper offers some recommendations.

Context-Setting: What are Automated Weapons Systems?

States have long used AI for military purposes, however, the use of AI-powered AWS is new.[2] Some AWS require a human to give an "initial command to attack" (human-in-the-loop) while others are fully autonomous and rely on data to determine when to deploy force.[3] In machine learning, "human-in-the-loop" refers to the need for human interaction or intervention to control an outcome.[4] Because AWS can be fully or semi-automated and have either lethal or non-lethal capabilities, consensus is lacking on a firm definition of what constitutes AWS; therefore, they are best viewed as a spectrum.[5]

This paper concerns those AWS that require little to no human involvement at the deployment stages, such as automated drones or "killer robots", that once activated by a human, use forms of AI to identify and engage targets.[6] While data about their deployment is limited, evidence suggests that AWS are being developed and tested by states including the U.S., China, and Russia and may have been used in Palestine, Ukraine, and Libya.[7]

---

[1] Sara Meger, "Che Guevara and the case for revolutionary feminism in global politics" (2022) 20:8 Globalizations 1581 at 1585; Judy Wajcman & Erin Young, "Feminism Confronts AI: The Gender Relations of Digitalization" in Jude Browne et al, *Feminist AI: Critical Perspectives on Algorithms, Data, and Intelligent Machines* (Oxford University Press, 2023) 47 at 52; I used ChatGPT to summarize critical feminist views of militarism and AI before conducting further research.

[2] Katherine Chandler, *Does Military AI Have Gender? Understanding bias and promoting ethical approaches in military applications of AI* (Geneva: UNIDIR, 2021) at 3.

[3] Birgitta Dresp-Langley, "The weaponization of artificial intelligence: What the public needs to be aware of" (2023) 6:1154184 Frontiers in Artificial Intelligence 1 at 2; United Nations Office for Disarmament Affairs, "Lethal Autonomous Weapon Systems (LAWS) – UNODA", (2023), online: <https://disarmament.unoda.org/the-convention-on-certain-conventional-weapons/background-on-laws-in-the-ccw/>

[4] Xiao-Li Meng, "Data Science and Engineering With Human in the Loop, Behind the Loop, and Above the Loop" (2023) 5:2 Harvard Data Science Rev 1 at 2.

[5] *Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects*, 19 April 2021, CCW/GGE.1/2020/WP.7 at 3, 7.

[6] Dresp-Langley, *supra* note 3 at 3; Neil Davison, "A legal perspective: Autonomous weapon systems under international humanitarian law" in United Nations Office of Disarmament Affairs Occasional Papers (Geneva: UN, 2018) 5 at 5; Tim McFarland, "Sciendo" (2024) 12:1 J Military Studies 75 at 76.

[7] Samuel Bendett, *The Role of AI in Russia's Confrontation with the West* (Washington: Centre for a New American Security, 2024) at 3, 4, 5; Lauren A Kahn, "Risky Incrementalism: Defense AI in the United States" in Heiko Borchert, Torben Schütz & Joseph Verbovszky, eds, *The Very Long Game: 25 Case Studies on the Global State of Defense AI* (Cham: Springer Nature Switzerland, 2024) 39 at 39, 40; Lauren Wilcox, "No Humans in the Loop: Killer Robots, Race, and AI" in Jude Browne et al, eds, *Feminist AI: Critical Perspecttives on Algorithms, Data, and Intelligent Machines* (Oxford: Oxford University Press, 2023) 83 at 85; Tejas Bharadwaj & Charukeshi Bhatt, "Understanding the Global Debate on Lethal Autonomous Weapons Systems: An Indian Perspective", (30 August 2024), online: <https://carnegieendowment.org/research/2024/08/understanding-the-global-debate-on-lethal-autonomous-weapons-systems-an-indian-perspective?lang=en>.

Discussion

Applying a Critical Feminist Lens (CFL) to AI and Militarism

Applying a CFL expands both the notion of what constitutes harm and who will be harmed by AWS, therefore revealing the devastating impacts of AWS that traditional assessments of AI and military decision-making neglect. In addition to employing gender perspectives, a CFL interrogates the "othering" and subsequent dehumanization of those who fall outside of Audre Lorde's "mythical norm," which describes the white, heterosexual, cisgender, able-bodied male.[8]

By assessing who is involved in making legal decisions or creating and training AI systems, a CFL reveals "the gender implications of rules and practices which might otherwise appear to be neutral or objective".[9] Gender gaps exist in both the AI and defense sectors; globally, women represent approximately 26% of data and AI workers, 13 per cent of senior military officers, and only 25 per cent of the workforce of the U.S.'s top five defence contractors.[10] Women are also underrepresented in the global discourse on arms control; at the first UN expert meeting on AWS, none of the eighteen experts invited were women.[11] These statistics are even less for racialized women.[12] Because it is largely those within the mythical norm that train AI systems, biases become ingrained in machine learning algorithms, thereby reproducing and exacerbating systemic inequalities based on race and gender.[13] This aligns with Wajcman's technofeminism theory, which views "technology as both a source and consequence of patriarchal relations".[14]

Militarism reinforces gendered and racialized hierarchies and othering through perpetuation of patriarchal values of power, subordination, and control; when these values are entrenched in technologies like AWS, which have the power to determine whose deaths are 'justifiable', patriarchy becomes automated.[15] The UN Group of Government Experts on Lethal Autonomous Weapons Systems (GGE LAWS) have acknowledged that AWS may perpetuate gender and racial biases, among others.[16]

AWS Will Decrease Adherence to IHL's Principles of Distinction and Proportionality

Armed conflict is guided by the principles of *jus ad bellum* ("the right to war"), which refers to the conditions under which states can use armed force, and *jus in bello* ("law in war"), which guides

---

[8] Audre Lorde, "Age, race, class, and sex: women redefining difference" in Maxine Baca Zinn et al, eds, *Gender Through the Prism of Difference*, 3rd ed (Oxford: Oxford University Press, 2005) 245 at 246.
[9] Katharine Bartlett, "Feminist Legal Methods" (1990) 103:4 Harvard L Rev 829 at 837.
[10] World Economic Forum, *Global Gender Gap Report 2020* (Geneva: World Economic Forum, 2020) at 38; United Nations, *Towards Equal Opportunity for Women in the Defence Sector* (New York: United Nations, 2024) at 5; Sarah Matar, Kanika Aggarwal & Jessica Groot, *Gender equality as a catalyst for aerospace and defense transformation*, by (Dubai: Kearny, 2024) at 2.
[11] Renata Hessmann Dalaqua, Kjølv Egeland & Torbjørn Graff Hugo, *Still Behind the Curve: Gender Balance in Arms Control, Non-proliferation and Disarmament*, (Geneva: UNIDIR, 2019) at 7; Sarah Knuckey, "Do women have anything to say about autonomous weapons?", (14 May 2014), online: <https://www.justsecurity.org/10424/women-autonomous-weapons/>.
[12] Wajcman & Young, *supra* note 1 at 49; Alexandra Stark & Heather Hurlburt, "Four Data Trends in Gender" (20 March 2020), online: <https://www.newamerica.org/political-reform/reports/four-data-trends-gender-diversity-and-security-you-should-know-about/>
[13] Wilcox, *supra* note 7 at 87; Wajcman & Young, *supra* note 1 at 48; Eleanor Drage & Federica Frabetti, "AI that Matters: A Feminist Approach to the Study of Intelligent Machines" in Jude Browne et al, eds, *Critical Perspectives on Algorithms, Data, and Intelligent Machines* (Oxford University Press, 2023) 274 at 274; Simon Lindgren, *Handbook of Critical Studies of Artificial Intelligence* (Cheltenham: Edward Elger Publishing, 2023) at 2.
[14] Wajcman & Young, *supra* note 1 at 52.
[15] Ray Acheson, *Autonomous Weapons and Patriarchy* (Geneva: Women's International League for Peace and Freedom, 2020) at 6, 17; Shimona Mohan, *Filling the Blanks: Putting Gender into Military A.I*, (New Delhi: Observer Research Foundation, 2023) at 4; Meger, *supra* note 1 at 1585.
[16] *Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons*, *supra* note 5 at 7.

how "moral" wars should be fought.[17] These concepts form the basis of the IHL principles of proportionality and distinction, which require "force be used discriminately to avoid harm to noncombatants and in proportion to the value of the military objective".[18] Although feminists argue that there is no justified legitimization of violence, IHL governs the use of weapons and therefore it is practical to assess AWS against these principles.[19]

Proponents of AWS argue that the increased accuracy and precision offered by AWS could reduce civilian casualties, thereby increasing adherence to distinction and proportionality.[20] However, there is consensus that AWS are not yet sophisticated enough (and may never be) to distinguish civilians from combatants.[21] Numerous studies have shown that because gender binaries are embedded in neural networks, commercial facial recognition technologies are inept at recognizing "non-binary, non-white subjects".[22] Further, the existing gendered norms of war, where combatants are presumed to be men and civilians women, mean that "algorithmic models might be less likely to identify civilian men as non-combatants".[23] As a whole, the use of identity characteristics as signifiers for drone strikes reinforces gender, racial, and ableist essentialisms.[24] Cultural differences might also influence AWS' misidentification of civilians. For example, predictions of "no civilian presence" prior to U.S. Government drone strikes in Iraq and Syria may have not accounted for the reality that "families were inside during Ramadan" - and this was with a human-in-the-loop.[25]

Militaries may embrace these algorithmic biases that categorize and identify to allow for intentional profiling based on race and gender or those who a state deems an "inherent risk" or an "other".[26] For example, the U.S. has long used automated drones as part of their counterterrorism projects to target "suspected terrorists" (Muslim and/or Arab men); the assistance of AI would optimize this racist targeting.[27] More recently, Israeli intelligence sources revealed that Israel uses an AI-powered database called Lavender to "mark suspected operatives of Hamas and Palestinian Islamic Jihad" and put them on kill lists.[28] Although Lavender makes errors in approximately ten percent of cases, and will "occasionally mark individuals who have merely a loose connection to militant groups, or no connection at all", minimal human assessment is used before approving a killing.[29]

---

[17] Alexander Orakhelashvili, *Akehurst's Modern Introduction to International Law*, 9th ed (New York: Routledge, 2022) at 481; ICRC, "Jus ad bellum and jus in bello", (28 July 2014), online: <https://www.icrc.org/en/law-and-policy/jus-ad-bellum-and-jus-bello>.

[18] Anthony Pfaff, "The Ethics of Acquiring Disruptive Military Technologies" (2020) 3 Texas National Security Rev 35 at 42; ICRC, *supra* note 17; *Additional Protocol I to the Geneva Conventions of 12 August 1949 and relating to the Protection of Victims of International Armed Conflicts*, 8 June 1977, arts 48, 51, 57.

[19] Meger, *supra* note 1 at 1584.

[20] McFarland, *supra* note 6 at 76, 78.

[21] Wilcox, *supra* note 7 at 86; *Report of the Special Rapporteur on extrajudicial, summary or arbitrary executions, Christof Heyns*, UNGA, 23rd Sess, UN Doc A/HRC/23/47 (2013).

[22] Drage & Frabetti, *supra* note 13 at 282, 286; Acheson, *supra* note 15 at 12; Lysanne Louter, "Racial bias in facial recognition algorithms", (21 March 2023), online: <https://www.amnesty.ca/features/racial-bias-in-facial-recognition-algorithms/>.

[23] Chandler, *supra* note 2 at 4.

[24] Drage & Frabetti, *supra* note 13 at 281, 282; Article 36 & Reaching Critical Will, *Sex and Drone Strikes: Gender and identity in targeting and casualty analysis* (London & New York: Article 36 & Reaching Critical Will, 2014) at 7.

[25] Wilcox, *supra* note 7 at 95.

[26] Acheson, *supra* note 15 at 12, 13, 14.

[27] John Rollins, *Armed Drones: Evolution as a Counterterrorism Tool* (Washington: Congressional Research Service, 2023) at 1, 2; Acheson, *supra* note 15 at 13.

[28] Amjad Iraqi, "'Lavender: The AI machine directing Israel's bombing spree in Gaza" (3 April 2024), online: <https://www.972mag.com/lavender-ai-israeli-army-gaza/>; Bethan McKernan & Harry Davies, "'The machine did it coldly': Israel used AI to identify 37,000 Hamas targets" (3 April 2024), online: <https://www.theguardian.com/world/2024/apr/03/israel-gaza-ai-database-hamas-airstrikes>.

[29] Iraqi, *supra* note 28.

Even if AWS could potentially reduce human casualties, this argument employs a narrow concept of harm. Applying feminist care ethics thickens the "thin concept of harm" that informs traditional military thinking about proportionality to include non-physical harm such as relational harm and psychological harm.[30] Proportionality assessments are highly contextual and often involve qualitative and therefore, human judgment.[31] Algorithms cannot conceptualize "the value of a human life" or non-physical harm and should therefore not be entrusted with the decision to choose who should live and die.[32] For example, when a drone strike spares some civilians from injury or death, *jus in bello* would find that "no injustice had occurred" because civilians who witnessed the strike did not experience physical harm.[33] A care ethics perspective would instead acknowledge the relational harm imposed on the surviving civilians through "the destruction of a relationship characterized by care and support".[34] AWS could also cause psychological harm; feminists argue that drone surveillance is a patriarchal tool that amounts to "psychological colonization".[35] For example, in Pakistan, a study revealed that the mere threat of being struck by U.S. drones caused psychological harm to some civilians.[36]

AWS Would Further Dehumanize Victims and Remove Combatants' Moral Agency

The concept of dehumanization is helpful in understanding the harms that AWS will likely have on the "the dehumanized inferior" - those who fall outside of the mythical norm.[37] When states view certain individuals as legally or socially "less than" or "illegible" as human, this is reflected in the laws and technologies that constitute their military apparatuses; consequently, human rights violations against those deemed inferior are not seen as true violations.[38] Israel dehumanizes Palestinians by framing them all as terrorists in an attempt to justify their breaches of IHL.[39] Because they view Palestinians as less than human, they are not incentivized to correct the errors of Lavender, for example. AWS are also seen by some as morally preferable, as they would reduce combatant casualties.[40] This perspective exalts the human worth of combatants, who are extensions of "state-legitimized violence" over the human worth of victims.[41]

While all military drones are "instruments of state violence", taking the human out of the loop removes the relational aspect between the drone and the target, which removes the target's humanity and the operator's moral agency.[42] When humans operate drones, they are well positioned to observe the "relationships among the people who inhabit prospective strike zones".[43] Observing the potential target's livelihood and relationships reveals the target's humanity, making

---

[30] Lindsay C. Clark & Christian Enemark, "Drone Warriors, Revealed Humanity and a Feminist Ethics of Care" in *Ethics Drone Strikes Restraining Remote Control Kill* (Edinburgh: Edinburgh University Press, 2021) 130 at 135.
[31] UNGA, *supra* note 21 at 14.
[32] Wilcox, *supra* note 7 at 93.
[33] Clark & Enemark, *supra* note 30 at 135.
[34] *Ibid* at 133, 135.
[35] Acheson, *supra* note 15 at 7; Anna Jackman & Katherine Brickell, "Everyday droning: Towards a feminist geopolitics of the drone-home" (2022) 46:1 Progress in Human Geography 156 at 159.
[36] Clark & Enemark, *supra* note 30 at 133.
[37] Lorde, *supra* note 8 at 245, 246; Catharine A. MacKinnon, *Are Women Human?: And Other International Dialogues* (Cambridge: Harvard University Press, 2006) 1 at 3.
[38] MacKinnon, *supra* note 37 at 3.
[39] Ahmad Ibsais, "Palestinians are being dehumanised to justify occupation and genocide", (20 August 2024), online:<https://www.aljazeera.com/opinions/2024/8/20/palestinians-are-being-dehumanised-to-justify-occupation-and-genocide>.
[40] Amitai Etzioni & Oren Etzioni, "Pros and Cons of Autonomous Weapons Systems" (2017) Military Rev (May-June 2017) at 72.
[41] Claire Duncanson & Rachel Woodward, "Regendering the military: Theorizing women's military participation" (2016) 47:1 Security Dialogue 3 at 5.
[42] Wilcox, *supra* note 7 at 87; Pfaff, *supra* note 18 at 44.
[43] Clark & Enemark, *supra* note 30 at 133.

dehumanization of the target more difficult.[44] Beyond the individual's own humanity, the drone operator may also become aware of how killing a target would adversely affect civilians who depend on the target's care, and therefore expand the operator's assessment of harm to account for relational harm.[45] In the early development of AI, feminist scholars posited that the AI field "was built on a model of intelligence that dissociated cognition from the body".[46] Removing the human-in-the-loop may cognitively and emotionally detach combatants from decisions to kill.[47] This dissociation, as well as the challenge of identifying a human to hold accountable for impermissible uses of force that AWS would likely create, may incentivize combatants to use force when not 'necessary'.[48]

AWS Will Harm Women Both in Conflict Zones and Beyond

AWS are a force multiplier that will likely lower "the threshold for conflict" and "lead to rapid conflict escalations and probable flash wars" thereby increasing both the instances and repercussions of armed conflict.[49] This will exacerbate the existing harms that armed conflict disproportionately imposes on women and girls, such as conflict-related sexual violence, disrupted access to reproductive healthcare, and the burden of relocating displaced families.[50]

Increased use of AWS would also harm those not in conflict zones; namely, poor, racialized women in the Global South. As demand increases for the raw materials needed to fuel the emerging military AI arms race, communities in the Global South bear the burden.[51] The majority of AI grunt work such as "data collection, cleaning, annotation, and algorithmic verification" is done by underpaid, exploited women in the Global South.[52] Further, the raw materials used in AI are largely mined and manufactured by "poor women from the Global South".[53] In the DRC, exploitive mining activities have resulted in human rights violations including sexual violence and vaginal infections due to contaminated water.[54]

Increased development and deployment of AWS would also worsen climate change, which already disproportionately burdens the Global South.[55] Armed conflict has an immense environmental impact; militaries account for approximately five and a half per cent of global GHG emissions and in both Ukraine and Gaza, there has been significant degradation of soil and contamination of water.[56] AI models are energy-intensive and their use in AWS will exacerbate climate change; mining causes environmental destruction, training AI models omits significant

[44] *Ibid*.
[45] *Ibid*.
[46] Chandler, *supra* note 2 at 9.
[47] UNGA, *supra* note 21 at 5.
[48] Pfaff, *supra* note 18 at 45.
[49] United Nations Office for Disarmament Affairs, *supra* note 3; Bharadwaj & Bhatt, *supra* note 7.
[50] Eran Bendavid et al, "The effects of armed conflict on the health of women and children" (2021) 397:10273 Lancet 522 at 5, 10, 11, 12, 13.
[51] Defense Advanced Research Projects Agency, "Increasing Transparency in Critical Materials Price, Supply, and Demand Forecasts", (25 October 2023), online: <https://www.darpa.mil/news-events/2023-10-24a>.
[52] Botlhokwa Ranta, *The Unknown Women of Content Moderation* [zine] (2024); Data Workers' Inquiry, "The hidden workers behind AI tell their stories" (8 July 2024), online: <https://netzpolitik.org/2024/data-workers-inquiry-the-hidden-workers-behind-ai-tell-their-stories/>.
[53] Emily Jones, "A Posthuman-Xenofeminist Analysis of the Discourse on Autonomous Weapons Systems and Other Killing Machines" (2018) 44:1 Australian Feminist L J 93 at 97.
[54] Amnesty International, *DRC: Powering Change or Business as Usual? Forced Evictions at Industrial Cobalt and Copper Mines in the Democratic Republic of Congo*, AFR 62/7009/2023 (London: Amnesty International, 2023) at 8, 13, 15, 21, 88.
[55] Intergovernmental Panel on Climate Change, *Climate Change 2022: Impacts, Adaptation and Vulnerability,* (Cambridge: Cambridge University Press, 2022) at 12, 13, 29.
[56] UN Peace and Security, "How conflict impacts our environment" (2023), online: <https://www.un.org/en/peace-and-security/how-conflict-impacts-our-environment>

carbon dioxide, and the rapid production of new AI technologies creates electronic waste which contaminates ecosystems.[57]

Further Reflections and Recommendations

While there is agreement in the international legal community that a treaty governing AWS is needed, one does not currently exist.[58] This leaves room for innovative approaches to the development and deployment of AWS. From a CFL perspective, a ban on AWS is the most preferable legal solution, as any weapons or technologies used for surveilling, controlling, or killing maintains existing power structures that promote white supremacy and patriarchy.[59] However, because AWS are already being tested, and potentially deployed, heavy regulation and feminist approaches to AWS are more realistic interventions.

For instance, the lack of critical interrogation of AI's impact on human rights in armed conflict is reflected in the Council of Europe's (CoE) international treaty addressing human rights and the rule of law in the use of AI systems.[60] While Article 4 requires parties to ensure that applications of AI are consistent with human rights obligations, Article 3(4) exempts parties from applying the treaty to systems related to the protection of its national defence.[61]

Ongoing efforts at the UN is not particularly discouraging. In November 2024, 165 states voted in favour of a UNGA resolution that mandates discussions about AWS, with a focus on ethical and human rights implications, and calls for negotiation of a legally binding treaty governing their use.[62] The UNGA also recently passed a draft resolution calling for the need to recognize women not only as victims of "gender-based armed violence" but as key players in arms control, disarmament, and non-proliferation policy making and practical implementation.[63] Responding to these calls by implementing feminist approaches to decision-making related to AWS at both national and international levels can interrogate how laws such as the CoE treaty neglect to truly account for human rights.

Moreover, the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons System recognizes that "achieving gender balance" in their group's work is important, but gender balance is not enough.[64] While increased representation is needed, it is not enough to view gender as a "soft security issue" or assess the legality and morality of AWS against traditional IHL principles; rather, international policy makers must employ feminist methods in assessing both the development and deployment of AWS immediately.[65] Tangible methods could include applying a tool such as Canada's Gender-based Analysis Plus to a potential treaty regarding AWS. This could reveal how their use would impact those from intersectional

---

[57] Alesia Zhuk, "Artificial Intelligence Impact on the Environment: Hidden Ecological Costs and Ethical-Legal Issues" (2023) 1:4 J Digital Technologies and L 932 at 936, 940.
[58] *Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law*, 5 May 2024, CETS 225 at 3, 4.
[59] Acheson, *supra* note 15 at 16, 17; A ban will remain the most preferable solution until it can be proven with certainty that AWS will consistently reduce human casualties.
[60] *Council of Europe*, *supra* note 58.
[61] *Ibid* at 3, 4.
[62] *General and complete disarmament: lethal autonomous weapons systems*, UNGA, 79th Sess, UN Doc A/C.1/79/L.43 (2024); United Nations, Press Release, GA/DIS/3757 "Fourteen New Drafts, Including on Implications of Artificial Intelligence in Military Domain, Approved in First Committee by 34 Votes" (6 November 2024), online: <https://press.un.org/en/2024/gadis3757.doc.htm>.
[63] *Women, disarmament, non-proliferation and arms control,* UNGA, 79th Sess, UN Doc A/C.1/79/L.69 (2024).
[64] *Ibid*; Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons System, *supra* note 5 at 7.
[65] Mohan, *supra* note 15 at 5.

identities and those not in conflict zones.[66] States like Canada and Mexico that have feminist foreign policies could leverage these existing legal principles to shape their development and deployment of AWS.[67]

AI used in AWS must also be overhauled and trained using technofeminist approaches. While starting with new AI models sounds utopian or impractical, there are methods that can minimize the biases embedded in AI. For AWS that use large language models, "machine unlearning" and data erasure practices can be applied to "refine the model's decision-making processes" to reduce biases and improve "interpretative accuracy".[68] Conducting performative analyses of AI to recognize how citations and agential cuts used at the training stages of AI lead to reproduction of gender and racial biases can allow those who train the AI in AWS to take a more informed approach that aligns with principles of feminist technoscience.[69] However, these strategies will not be effective unless international law requires states to use them.

Conclusion

AWS have existential implications for humanity. International law makers must act now to regulate the development and deployment of AWS in alignment with critical feminist perspectives. As AWS become commonplace in armed conflict, feminist-informed research should be conducted to determine the impact of AWS on the human rights of individuals in conflict zones.

[66] Women and Gender Equality Canada, "What is Gender-based Analysis Plus", (2 May 2024), online: <https://www.canada.ca/en/women-gender-equality/gender-based-analysis-plus/what-gender-based-analysis-plus.html>.
[67] Johanna Nelles, "A trend that's catching on? Feminist foreign policy and international efforts to end violence against women", (8 March 2023), online: <https://www.mcgill.ca/humanrights/article/trend-thats-catching-feminist-foreign-policy-and-international-efforts-end-violence-against-women>; Global Affairs Canada, *Canada's Feminist International Assistance Policy*, (Ottawa: Global Affairs Canada, 2017).
[68] Youyang Qu et al, *The Frontier of Data Erasure: Machine Unlearning for Large Language Models* (2024) arXiv:2403.15779 at 1.
[69] Drage & Frabetti, *supra* note 13 at 277, 279, 285, 286.

References

Legislation

Additional Protocol I to the Geneva Conventions of 12 August 1949 and Relating to the Protection of Victims of International Armed Conflicts, 8 June 1977.

Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, 19 April 2021, CCW/GGE.1/2020/WP.7.

Council of Europe, *Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law*, 5 May 2024, CETS 225.

General and Complete Disarmament: Lethal Autonomous Weapons Systems, UNGA, 79th Sess, UN Doc A/C.1/79/L.77 (2024).

Women, Disarmament, Non-Proliferation and Arms Control, UNGA, 79th Sess, UN Doc A/C.1/79/L.69 (2024).

Books

Davison, Neil, "A Legal Perspective: Autonomous Weapon Systems Under International Humanitarian Law" in *United Nations Office of Disarmament Affairs Occasional Papers* (Geneva: United Nations, 2018) 5.

Drage, Eleanor & Federica Frabetti, "AI that Matters: A Feminist Approach to the Study of Intelligent Machines" in Jude Browne et al, eds, *Critical Perspectives on Algorithms, Data, and Intelligent Machines* (Oxford: Oxford University Press, 2023) 274.

Lindgren, Simon, *Handbook of Critical Studies of Artificial Intelligence* (Cheltenham: Edward Elger Publishing, 2023).

Lorde, Audre, "Age, Race, Class, and Sex: Women Redefining Difference" in Maxine Baca Zinn et al, eds, *Gender Through the Prism of Difference*, 3rd ed (Oxford: Oxford University Press, 2005) 245.

MacKinnon, Catharine A., *Are Women Human?: And Other International Dialogues* (Cambridge: Harvard University Press, 2006).

Mohan, Shimona, *Filling the Blanks: Putting Gender into Military A.I* (New Delhi: Observer Research Foundation, 2023).

Wilcox, Lauren, "No Humans in the Loop: Killer Robots, Race, and AI" in Jude Browne et al, eds, *Feminist AI: Critical Perspectives on Algorithms, Data, and Intelligent Machines* (Oxford: Oxford University Press, 2023) 83.

Wajcman, Judy & Erin Young, "Feminism Confronts AI: The Gender Relations of Digitalisation" in Jude Browne et al, eds, *Feminist AI: Critical Perspectives on Algorithms, Data, and Intelligent Machines* (Oxford; Oxford University Press, 2023) 47.

Journal Articles

Bartlett, Katharine, "Feminist Legal Methods" (1990) 103:4 Harvard Law Review 829.

Bendavid, Eran, et al., "The Effects of Armed Conflict on the Health of Women and Children" (2021) 397:10273 Lancet 522.

Duncanson, Claire & Rachel Woodward, "Regendering the Military: Theorizing Women's Military Participation" (2016) 47:1 Security Dialogue 3.

Dresp-Langley, Birgitta, "The Weaponization of Artificial Intelligence: What the Public Needs to Be Aware of" (2023) 6:1154184 Frontiers in Artificial Intelligence 1.

Jackman, Anna & Katherine Brickell, "Everyday Droning: Towards a Feminist Geopolitics of the Drone-Home" (2022) 46:1 Progress in Human Geography 156.

Jones, Emily, *A Posthuman-Xenofeminist Analysis of the Discourse on Autonomous Weapons Systems and Other Killing Machines* (2018) 44:1 Australian Feminist L J 93.

Meng, Xiao-Li, "Data Science and Engineering With Human in the Loop, Behind the Loop, and Above the Loop" (2023) 5:2 Harvard Data Science Rev 1.

Pfaff, Anthony, "The Ethics of Acquiring Disruptive Military Technologies" (2020) 3 Texas National Security Rev 35.

Zhuk, Alesia, "Artificial Intelligence Impact on the Environment: Hidden Ecological Costs and Ethical-Legal Issues" (2023) 1:4 J Digital Technologies and L 932.

Reports

Acheson, Ray, *Autonomous Weapons and Patriarchy* (Geneva: Women's International League for Peace and Freedom, 2020).

Amnesty International, *DRC: Powering Change or Business as Usual? Forced Evictions at Industrial Cobalt and Copper Mines in the Democratic Republic of Congo*, AFR 62/7009/2023 (London: Amnesty International, 2023).

Chandler, Katherine, *Does Military AI Have Gender? Understanding Bias and Promoting Ethical Approaches in Military Applications of AI* (Geneva: UNIDIR, 2021).

United Nations, *Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions*, Christof Heyns, UNGA, 23rd Sess, UN Doc A/HRC/23/47 (2013).

United Nations, *Towards Equal Opportunity for Women in the Defence Sector* (New York: United Nations, 2024).

World Economic Forum, *Global Gender Gap Report 2020* (Geneva: World Economic Forum, 2020).

Online Materials

Bharadwaj, Tejas & Charukeshi Bhatt, "Understanding the Global Debate on Lethal Autonomous Weapons Systems: An Indian Perspective" (30 August 2024), online: <https://carnegieendowment.org/research/2024/08/understanding-the-global-debate-on-lethal-autonomous-weapons-systems-an-indian-perspective?lang=en>.

Iraqi, Amjad, "'Lavender: The AI Machine Directing Israel's Bombing Spree in Gaza" (3 April 2024), online: <https://www.972mag.com/lavender-ai-israeli-army-gaza/>

Knuckey, Sara, "Do Women Have Anything to Say About Autonomous Weapons?" (14 May 2014), online: <https://www.justsecurity.org/10424/women-autonomous-weapons/>.

McKernan, Bethan & Harry Davies, "'The Machine Did It Coldly': Israel Used AI to Identify 37,000 Hamas Targets" (3 April 2024), online: <https://www.theguardian.com/world/2024/apr/03/israel-gaza-ai-database-hamas-airstrikes>.

Ranta, Botlhokwa, *The Unknown Women of Content Moderation* [zine] (2024), online: <https://data-workers.org/wp-content/uploads/2024/06/Proj-Ranta-Version-v8-final.pdf>.

Stark, Alexandra & Heather Hurlburt, "Four Data Trends in Gender" (20 March 2020), online: <https://www.newamerica.org/political-reform/reports/four-data-trends-gender-diversity-and-security-you-should-know-about/>

United Nations, Press Release, GA/DIS/3757 "Fourteen New Drafts, Including on Implications of Artificial Intelligence in Military Domain, Approved in First Committee by 34 Votes" (6 November 2024), online: <https://press.un.org/en/2024/gadis3757.doc.htm>.

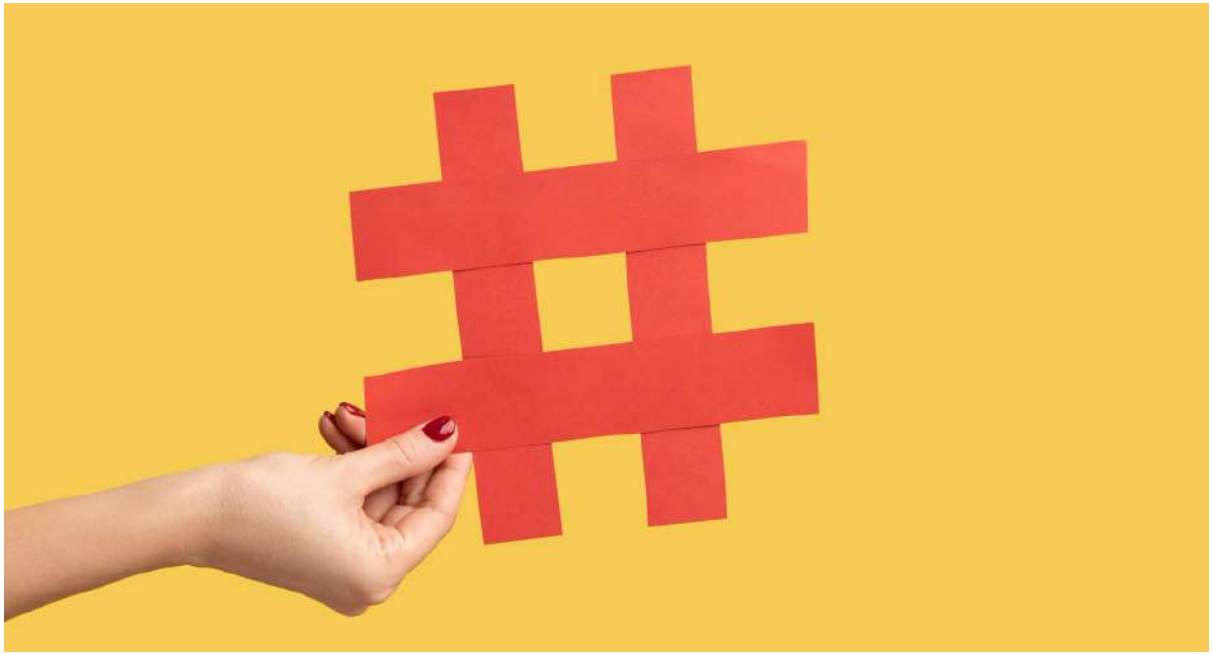
United Nations Office for Disarmament Affairs, "Lethal Autonomous Weapon Systems" (2023), online:<https://disarmament.unoda.org/the-convention-on-certain-conventional-weapons/background-on-laws-in-the-ccw/>

Chapter 11

Algorithmic Moderation and Foreign Interference: The Case of Sikh Activism On Social Media

Sunny Pabla (He/Him/Il)

JD Candidate, Lincoln Alexander School of Law



Abstract

This paper explores how social media platforms, driven by algorithms, inform popular discourse and shape public opinion. It focuses on the impact of compromised algorithms on vulnerable groups - in particular, the suppression of Sikh activists' voices by the Indian State. By suppressing Sikh voices and promoting counter-narratives that paint Sikh activists as 'extremists' and 'radicals', the Indian State is able to exploit the algorithm and shape the narrative on Sikh activists and activism. Drawing on critical theories of algorithmic bias and foreign interference, it positions the problem within the broader context of data, AI, and the law, questioning the ethical and legal responsibilities of social media companies in curating content and addressing potential governmental influence. It takes the stance that the current regulatory frameworks are inadequate to protect minority voices from suppression, resulting in systemic distrust in digital spaces. Overall, the paper summarily proposes possible legal interventions focused on increasing transparency in algorithmic moderation, enforcing accountability measures for platform manipulation driven by state actors, and establishing an independent oversight body to monitor such interference. While appreciating the challenges arising with moderation by social media platform, the study advocates for regulatory reforms to protect political expression, particularly for marginalized voices, and to strengthen democratic integrity on global social media platforms.

Keywords: Algorithmic Bias, Social Media Regulation, Foreign Interference, Pro-Sikh Activism, Content Moderation.

Introduction

This paper explores praxis of suppression in the context of pro-Sikh activism on social media. As the case studies would demonstrate, algorithmic vulnerabilities and state directed algorithmic governance can undermine democratic discourse. To further contextualize the analytical thrust of the paper, the paper draws on critical theories of algorithmic bias, freedom of speech, and foreign interference to analyze the suppression of pro-Sikh activism.

Algorithmic bias occurs when systems produce inequitable outcomes, often reflecting prejudices embedded in training data. Algorithmic biases unintentionally perpetuate societal inequities, as seen in platforms disproportionately flagging pro-Sikh content as extremist.[1] As Jasjit Singh research noted, justice-oriented activism is targeted by algorithmic moderation.[2] Moreover, algorithmic practices like "shadow banning" and opacity in algorithmic decision-making undermine freedom of speech and silent dissent voices - is vital feature in democratic societies.[3]

With States' interests added (in this case, the Indian State), algorithmic vulnerabilities can be further manipulated to systematically choke dissenting narratives.[4] Indeed, algorithmic moderation, as a modern form of institutional control, perpetuates this silencing by flagging activist content as extremist or prioritizing state-backed disinformation. These practices exacerbate the challenges of balancing moderation with preserving free expression and emphasize the urgent need for accountability and transparency in content governance. Already, reports from Reporters Without Borders[5] and Al Jazeera[6] highlight India's use of regulatory pressure to remove critical content while amplifying pro-government narratives. These actions underscore the global implications of unchecked state influence on platform governance.

What these perspectives illustrates quite evidently is that though social media platforms can amplify marginalized voices, they are not neutral; their algorithmic systems, shaped by inherent susceptibilities and external-cum-state-driven influences, are often manipulated to suppress specific voices under the guise of engagement or neutrality. Algorithmic moderation can amplify biases, marginalize communities, and silence dissent, particularly when state actors manipulate and exploit these systems to stifle activism. The weaponization of artificial intelligence in geopolitical conflicts has emerged as a critical concern, demonstrating how these tools can be leveraged to suppress political discourse and advocacy globally.

Context and Systemic Challenges – Pro-Sikh Advocacy

The Sikh community has long experienced systemic challenges. Indian interference highlights a troubling trend where Sikh activists advocating for human rights and Khalistan, an independent Sikh homeland, are labelled as extremists. This manipulation of digital platforms highlights

[1] Emily Bembeneck, Rebecca Nissan & Ziad Obermeyer, "To Stop Algorithmic Bias, We First Have to Define It" (21 October 2021), online: Brookings Institution https://www.brookings.edu/articles/to-stop-algorithmic-bias-we-first-have-to-define-it/.
[2] Jasjit Singh, "Narratives in Action: Modelling the Types and Drivers of Sikh Activism in Diaspora" (2020) 11:10 *Religions* 539.
[3] John Stuart Mill, *On Liberty* (London: John W Parker & Son, 1859).
[4] See the work of Noam Chomsky, *The Responsibility of Intellectuals* (New York: New York Review of Books, 1967) who critiqued the mechanisms through which dissent is systematically silenced. Chomsky argues that power structures manipulate narratives to delegitimize opposition, not always through overt censorship but often by creating environments where alternative viewpoints are marginalized or ignored.
[5]Reporters Without Borders, "Modi Ramps Up Online Censorship in India" (4 September 2023), online: https://rsf.org/en/modi-ramps-online-censorship-india.
[6] Al Jazeera, "In India, Government Now Has Power Over Social Media Content Moderation" (28 October 2022), online: https://www.aljazeera.com/economy/2022/10/28/in-india-govt-now-has-power-over-social-media-content-moderation.

broader concerns about state actors exploiting algorithmic systems to suppress dissent and assert control.[7]

Historical events like Operation Blue Star in 1984, when the Indian Army stormed Harmandir Sahib, a central Sikh place of worship, continue to resonate with the Sikh diaspora. Viewed as an attack on religious identity, this event catalyzed global activism and a renewed focus on Sikh sovereignty.[8] These historical grievances are amplified in the digital age, where platforms suppress dissent while amplifying state-sponsored narratives. Further exacerbating this issue, intelligence reports from the 2021 Canadian federal election indicate that India clandestinely provided financial support to preferred candidates, allegedly to sway narratives in its favor. Such activities exemplify how foreign interference intersects with algorithmic biases, where disinformation campaigns targeting Sikh advocacy are amplified while genuine activism is suppressed.[9]

Since 2020, hundreds of Sikhs have been detained for online activities, such as sharing posts about Khalistan, often based on algorithmically flagged content lacking substantive evidence.[10] Many detainees have faced severe mistreatment, including custodial sexual violence, as exemplified by the case of Nodeep Kaur. Kaur, a young Dalit labour rights activist, was arrested during the Farmers' Protests and faced physical and sexual abuse while in custody. Her case drew international attention to the use of excessive force and systemic targeting of activists, both online and offline.[11]

During the Farmers' Protests, written testimony on disinformation by the *Sikh American Legal Defense and Education Fund* ("SALDEF") revealed that hashtags like "#Sikh" and "#Sikhism" were systematically suppressed, while pro-government narratives were amplified.[12] This pattern of digital suppression aligns with allegations of India's broader interference tactics, including efforts to influence Canadian electoral outcomes through proxies and disinformation campaigns. Such activities are not confined to Indian borders but extend to platforms and narratives globally, disproportionately affecting the Sikh diaspora.[13] An article from Baaz News by Jasmeen Bassi highlights that "Sikh ground reporters and kisan mazdoor ekta movement affiliated accounts such as @iamparmjit, @sikhsiyasat, @PunYaab, @panth_punjab, @Kisanektamorcha, @Tractor2twitr, and @kisaanivichaar were suspended by Twitter en par with what seems to be a growing pattern of Sikh censorship on social media."[14] These patterns highlight the role of algorithmic bias in marginalizing vulnerable communities.

[7] Office of the United Nations High Commissioner for Human Rights, "Business and Human Rights" (2024), online: OHCHR https://www.ohchr.org/en/business-and-human-rights.
[8] Jasjit Singh, "Racialisation, 'Religious Violence' and Radicalisation: The Persistence of Narratives of 'Sikh Extremism'" (2020) 46:15 *Journal of Ethnic and Migration Studies* 3136 at 3147.
[9] Public Inquiry Into Foreign Interference in Federal Electoral Processes and Democratic Institutions, Final Report, vol 1: Report Summary (Toronto: Privy Council Office, 2025) at 40.
[10] World Sikh Organization of Canada, *Enforcing Silence: India's War on Sikh Social Media* (July 2020).
[11] BBC News, "Nodeep Kaur: Indian Activist Allegedly Beaten and Sexually Assaulted in Custody" (15 February 2021), online: BBC https://www.bbc.com/news/world-asia-india-56071706.
[12] Sikh American Legal Defense and Education Fund, "Written Testimony on Disinformation Nation: Social Media's Role in Promoting Extremist and Misinformation," submitted to the United States House of Representatives, Committee on Energy & Commerce, Subcommittee on Communications and Technology and Subcommittee on Consumer Protection and Commerce (25 March 2021).
[13] Supra note 9.
[14] Jasmeen Bassi, "How India Uses Communications to Suppress Farmers and Sikhs" (2021), Baaz News online: https://www.baaznews.org/p/jasmeen-bassi-how-india-uses-communications?utm_source=publication-search.

Reports from AP News[15] and Le Monde[16] reveal how India pressured platforms to silence Sikh voices globally. This interference aligns with a broader pattern of censorship and suppression targeting Sikh activists and organizations. Rupi Kaur, a globally recognized poet and activist, criticized Twitter's role in this suppression during the Farmers' Protests in India. On January 26, 2021, Kaur highlighted the suspension of accounts documenting the protests, including @SikhSiyasat, @Kisanektamorcha, and @IamParmjit, and questioned Twitter's complicity in amplifying state-backed disinformation while silencing dissent. These suspensions occurred amid an internet ban in India, further restricting activists' ability to document and engage with ongoing protests.[17]

Recent reports also shed light on how India's censorship tactics transcend its borders. For example, a Canadian Sikh advocacy group; The World Sikh Organization ("WSO"), accused Twitter of censoring its posts at the request of the Indian government. According to the National Post, these takedowns reveal how India's influence over global tech platforms threatens the ability of diaspora communities to freely advocate for human rights and political causes.[18] Similarly, CBC News reported on how Twitter blocked content from Canadian Sikh organizations and public figures, including tweets from Canadian poet and author, Rupi Kaur and the leader of the New Democratic Party of Canada, Jagmeet Singh, in compliance with Indian government requests.[19] The targeted removal of posts, such as those amplifying dissent or criticizing state policies, highlights the extraterritorial reach of state-sponsored censorship. This extends even to nations like Canada, which have stronger protections for free expression. Such actions raise significant concerns about the complicity of platforms in suppressing freedom of expression, even in jurisdictions where these rights are constitutionally enshrined. The 2023 assassination of Canadian Sikh activist Hardeep Singh Nijjar further highlights the extent of state interference.

Since 2020, numerous Sikh activists and media accounts have faced similar treatment, with many suspended or censored under the guise of legal compliance. Jas UK Singh, a prominent advocate, received an official notice from Twitter about the withholding of his account in accordance with India's Information Technology Act, 2000.[20] This exemplifies how platforms enforce local government requests to suppress dissenting voices. Furthermore, hashtags like #FreeJaggiNow and #NeverForget84, which amplify calls for justice and remembrance of significant Sikh historical events, were also censored, illustrating the systematic targeting of Sikh advocacy online.

Kaur's critique stresses the disproportionate impact of algorithmic moderation and content takedowns on Sikh activists, raising significant concerns about platform accountability and state

[15] AP News, "Canada-India Diplomatic Tensions: Allegations of Violence and State-Sponsored Actions" (September 2023), online: https://apnews.com/article/canada-india-diplomats-violence-80e9a0d43faa7db99781563434d0f1e0.
[16] Le Monde, "Canada Accuses India of Criminal Activity on Its Soil" (October 2024), online: https://www.lemonde.fr/en/international/article/2024/10/17/canada-accuses-india-of-criminal-activity-on-its-soil_6729631_4.html.
[17] Rupi Kaur, "@Twitter has censored Sikh media accounts which have been showcasing the reality of protestors on the ground... this decision—amid an internet ban currently enforced by India—will cost lives. #FarmersProtest #SikhCensorship" (26 January 2021, 3:28 PM), online: Twitter https://x.com/rupikaur_/status/1354164407329153024.
[18] National Post, "Canadian Sikh Group Alleges Censorship After Indian Government Asks Twitter to Delete Its Post" (2023), online: National Post https://nationalpost.com/news/canada/canadian-sikh-group-alleges-censorship-after-indian-government-asks-twitter-to-delete-its-post.
[19] CBC News, "Twitter Blocked Canadian Tweets at India's Request, Including Those by Rupi Kaur and Jagmeet Singh" (2023), online: CBC News https://www.cbc.ca/news/canada/british-columbia/twitter-canada-india-rupi-kaur-jagmeet-singh-1.6787760.
[20] Jas UK Singh, "This is the reward for speaking the truth, raising human rights and challenging oppression & inequality... #Censorship Well only in India! @jack @Twitter @TwitterIndia #FreeJaggiNow #NeverForget84 #SikhGenocide84 #Sikhs #Truth #Democracy #FreedomOfSpeech #UndueInfluence #Silencing" (18 February 2022, 5:08 PM), online: Twitter https://x.com/UK51NGH/status/1494795952011554819

influence over global tech companies.[21] These actions not only stifle legitimate activism but also amplify broader questions about sovereignty and the urgent need for international governance to address foreign interference in digital spaces.

Hardeep Singh Nijjar's Advocacy – Algorithmic Manipulation

Nijjar experienced targeted suppression online, where Indian agents reportedly exploited algorithms to brand his advocacy as extremist. A WSO report states that this tactic reflects a broader trend of criminalizing dissent through algorithmic bias.[22] Journalist Rana Ayyub has also highlighted how state mechanisms in India systematically label dissenting voices, including activists and journalists, as extremist or anti-national to justify their suppression. Her analysis highlights the broader convergence of algorithmic bias and state-led efforts to marginalize advocacy and stifle political discourse.[23] This aligns with global digital authoritarianism, where state actors manipulate algorithms on various platforms to silence political discourse.

Jaskaran Sandhu, in his article "India is Now the World's Largest Electoral Autocracy," delves into how legal frameworks like the *Unlawful Activities (Prevention) Act* ("UAPA") have been systematically weaponized to silence dissent. He highlights that the UAPA's broad definitions and provisions enable the Indian government to arbitrarily designate individuals as terrorists, bypassing judicial processes and criminalizing legitimate activism.[24] This legal framework complements digital strategies like algorithmic bias in targeting activists like Nijjar, showcasing the intersection of state law and technology in suppressing political dissent.

In the recently released Public Inquiry into Foreign Interference in Federal Electoral Processes and Democratic Institutions report (January 28, 2025), The Honourable Marie-Josée Hogue stated:

> "India also uses disinformation as a key form of foreign interference against Canada, a tactic likely to be used more often in the future. Until recently, Canada was trying to improve its bilateral relationship with India. However, the assassination of Hardeep Singh Nijjar, coupled with credible allegations of a potential link between agents of the Government of India and Mr. Nijjar's death, derailed those efforts. India has repeatedly denied these allegations. In October 2024, Canada expelled six Indian diplomats and consular officials in reaction to a targeted campaign against Canadian citizens by agents linked to the Government of India."

This statement emphasizes the growing role of disinformation as a strategic tool of foreign interference, particularly in the context of deteriorating diplomatic relations. The same report further reveals India's influence extending beyond its borders, including allegations of clandestine financial support for preferred candidates during Canada's 2021 federal election. These actions point to a coordinated effort to manipulate both digital and political landscapes in tandem. By shaping electoral narratives while leveraging algorithmic tools to suppress dissent (such as Nijjar's advocacy), India exemplifies the intersection of foreign interference tactics and digital suppression, which collectively undermine democratic principles on a global scale.[25] This convergence highlights the pressing need for robust policies that safeguard activism, promote algorithmic

[21] Supra note 17
[22] Supra note 10.
[23] World Sikh Organization, "Criminalizing Dissent in India: A Conversation with Rana Ayyub," YouTube (28 October 2021), online: https://www.youtube.com/watch?v=FUQ8ls_Flb4.
[24] Jaskaran Sandhu, "India Is Now The World's Largest Electoral Autocracy," Baaz News (22 August 2023), online: https://www.baaznews.org/p/jaskaran-sandhu-india-is-now-the?utm_source=publication-search.
[25] Supra note 9.

transparency, and resist undue influence from state actors seeking to exploit digital platforms for geopolitical advantage.

The Role of Bot Networks & Platform Inconsistencies

Studies have revealed bot networks linked to Indian actors that flood social media with anti-Sikh propaganda, using hashtags like #RealSikhsAgainstKhalistanis to delegitimize activism during the Farmers' Protests.[26] These campaigns amplify false narratives, drowning out legitimate discourse and demonstrating how AI tools can be weaponized in geopolitical conflicts. This emphasizes the urgent need for algorithmic audits and public accountability.[27]

India's activities primarily target Canada's Sikh diaspora of approximately 800,000 people, promoting a pro-India and anti-Khalistan narrative. According to the Public Inquiry Into Foreign Interference report, these actions are consistent with classified evidence linking violent criminal activity, including homicides and extortion, to agents of the Indian government. Furthermore, the report identifies India as an emerging cyber threat actor, underscoring the sophisticated digital tactics employed to suppress dissent. Bot networks serve as a key component of this strategy, flooding platforms with disinformation to delegitimize Sikh activism while amplifying propaganda. This digital repression silences legitimate voices within the diaspora and highlights the intersection of algorithmic exploitation and state-sponsored transnational repression. It highlights the critical need for enhanced accountability measures to mitigate the misuse of AI-driven tools in geopolitical conflicts.[28]

Moreover, social media platforms often inconsistently enforce moderation policies, disproportionately flagging pro-Sikh content while permitting harmful propaganda.[29] During the Farmers' Protests, platforms like Twitter and Instagram blocked Sikh-related hashtags but allowed inflammatory ones like #Shoot to proliferate. Such disparities erode trust and disproportionately harm marginalized communities.[30]

These challenges mirror global examples. In Myanmar, Facebook's algorithms amplified content inciting violence against the Rohingya minority, contributing to atrocities.[31] Similarly, in Russia, digital platforms suppressed LGBTQ+ activism under state pressure.[32] These cases highlight algorithmic failures and reinforce the need for transparency and accountability in digital governance.

Real-time content moderation remains a significant challenge. Platforms struggle to manage live-streamed material during protests, allowing harmful content to spread unchecked while over-

---

[26] Centre for Information Resilience, "Return of the RealSikhs: The Fake Network Targeting Sikhs Across the World Despite Platform Takedowns" (2023), online: CIR https://www.info-res.org/cir/articles/return-of-the-realsikhs-the-fake-network-targeting-sikhs-across-the-world-despite-platform-takedowns/.
[27] Geo News, "Indian Fake Media Network Targeting Sikhs, Farmers, Separatists Exposed in UK Research" (2021), online: Geo News https://www.geo.tv/latest/384101-indian-fake-media-network-targeting-sikhs-farmers-separatists-exposed-in-uk-research.
[28] Supra note 9.
[29] The Washington Post, "India's Facebook Problem: Propaganda and Hate Speech Flourish Ahead of Elections" (26 September 2023), online: The Washington Post https://www.washingtonpost.com/world/2023/09/26/india-facebook-propaganda-hate-speech/.
[30] Supra note 12.
[31] Amnesty International, "Myanmar: Facebook's Systems Promoted Violence Against Rohingya; Meta Owes Reparations – New Report" (29 September 2022), online: Amnesty International https://www.amnesty.org/en/latest/news/2022/09/myanmar-facebooks-systems-promoted-violence-against-rohingya-meta-owes-reparations/.
[32] SpringerLink, "Virtually (Non)existent? The Role of Digital Media in Russian LGBTQ+ Activism" (July 2023), online: SpringerLink https://link.springer.com/article/10.1057/s41311-024-00592-7.

censoring legitimate discourse. Measures like delayed streaming and real-time oversight teams are critical to mitigating these issues.[33]

Ethical And Legal Concerns

a. Algorithmic Bias and Inequality

Algorithms are not neutral; they reflect biases embedded in their training data. Terms like "Khalistan" are often flagged as extremist, silencing legitimate voices while leaving harmful content unaddressed.[34] Shoshana Zuboff's *The Age of Surveillance Capitalism* highlights how platforms amplify state narratives while neglecting minority protections.[35] A parallel can be drawn to the Ethiopian conflict, where Facebook's algorithms failed to curb hate speech and incitement to violence against ethnic Tigrayans. Reports revealed that inflammatory posts, many of which were left unchecked, contributed to an environment of hostility and violence, underlining the detrimental role of algorithmic bias in exacerbating real-world harm.[36]

Algorithms frequently lack contextual understanding, leading to over-censorship. For example, another instance of this would include Tumblr's 2018 content ban mistakenly flagged benign material, illustrating the limitations of automated moderation without human oversight.[37] Such failures are especially damaging in politically sensitive contexts like Sikh activism.[38]

b. Transparency and Accountability

The lack of transparency in content moderation practices allows platforms to perpetuate biases and censorship. The *United Nations Guiding Principles on Business and Human Rights* ("UNGPs") emphasize corporate accountability in preventing complicity in human rights violations.[39] However, platforms fail to address systemic gaps. For example, Meta[40] has been accused of allowing harmful content ahead of Indian elections while disproportionately removing Sikh-related posts.[41]

Governments exacerbate these issues by pressuring platforms. Jack Dorsey revealed that India threatened to shut down Twitter during the Farmers' Protests unless critical content was removed. Such incidents highlight the tension platforms face between upholding global free speech norms and complying with state demands.[42]

---

[33] MacKenzie F. Common, "Fear the Reaper: How Content Moderation Rules Are Enforced on Social Media" (2020) 34:2 Int'l Rev of L, Computers & Tech 126.
[34] Supra note 10.
[35] Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (2019).
[36] "Meta's Failure Contributed to Abuses Against Tigray in Ethiopia" (10 October 2023), online: Amnesty International https://www.amnesty.org/en/latest/news/2023/10/meta-failure-contributed-to-abuses-against-tigray-ethiopia/.
[37] Greyson K. Young, "How Much Is Too Much: The Difficulties of Social Media Content Moderation" (2022) 31:1 Information & Communications Technology Law
[38] Supra note 33 at 126.
[39] Office of the High Commissioner for Human Rights, *Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework* (Geneva: United Nations, 2011), online: OHCHR https://www.ohchr.org/en/publications/reference-publications/guiding-principles-business-and-human-rights.
[40] Business & Human Rights Resource Centre, "Meta responds to allegations of content moderation failures facilitating misinformation, harmful content ahead of Indian elections" (2023), online: https://www.business-humanrights.org/en/latest-news/meta-responds-to-allegations-of-content-moderation-failures-facilitating-misinformation-harmful-content-ahead-of-indian-elections/.
[41] The Sikh Coalition, "Holding Social Media Platforms Accountable for Censorship" (2024), online: https://www.sikhcoalition.org/our-work/legal-and-policy/holding-social-media-platforms-accountable-for-censorship/.
[42] Independent, "Jack Dorsey Says Twitter Was Pressured by India During Farmers' Protests" (2023), online: Independent https://www.independent.co.uk/tech/jack-dorsey-twitter-india-protests-b2356403.html.

c. Freedom of Speech

The suppression of pro-Sikh activism undermines fundamental freedoms of speech and expression. The U.S. State Department's 2023 Human Rights Report highlights India's use of platform censorship to stifle dissent.[43] Diaspora nationalism, a vital form of cultural preservation, is often mischaracterized as extremism. This further marginalizes these voices.[44] The Sikh Coalition's 2024 report further emphasizes how Indian authorities have weaponized digital platforms to silence dissent, illustrating the broader implications of algorithmic moderation that fails to distinguish legitimate advocacy from extremism.[45]

Additionally, the suppression of pro-Sikh activism on social media raises critical concerns under international human rights law, particularly regarding freedom of expression and political participation. Article 19 of the *International Covenant on Civil and Political Rights* ("ICCPR") protects the right to hold opinions without interference and to seek, receive, and impart information freely. Algorithmic moderation that silences minority voices, coupled with state interference, directly violates these rights.[46]

Article 21 of the ICCPR guarantees the right to peaceful assembly, a principle increasingly relevant as digital platforms become central to political discourse and mobilization. Restrictions on online activism undermine these rights, and states have a duty under international law to ensure private actors, such as social media companies, do not infringe upon them.[47]

The UNGPs emphasize corporate responsibility in preventing human rights violations. Social media companies must implement due diligence processes to identify and mitigate abuses, yet reliance on opaque algorithmic systems undermines these obligations. Stronger regulatory measures are urgently needed.[48]

Proposed Legal Interventions

Algorithmic moderation and foreign interference present significant concerns, and understandably so that social media platforms face inherent challenges in balancing free speech with content moderation responsibilities. These challenges include the sheer volume of daily contents that require moderation, multiplicity of legal regulation that present compliance complexities, government pressure vis the market imperatives for social media companies to maintain market share, and the possibility of a chaotic and hate explosive social media space due to relax moderation rules and policies. Truly, this limitation exacerbates the over-censorship of marginalized voices while allowing harmful content to proliferate. Yet, deliberate government censorship of dissent voices should not be accommodated under any circumstances where the group concerned as in this case seek legitimate interests.

---

[43] United States, Department of State, *2023 Country Reports on Human Rights Practices: India* (2023), online: Department of State https://www.state.gov/reports/2023-country-reports-on-human-rights-practices/india/.
[44] Supra note 12.
[45] Sikh Coalition, *So Many Targets: Contextualizing Modern Indian Transnational Repression Against the Sikh Community* (2024), online: Sikh Coalition https://www.sikhcoalition.org/wp-content/uploads/2024/07/So-Many-Targets-Sikh-Coalition-TNR-Report.pdf.
[46] International Covenant on Civil and Political Rights, 16 December 1966, 999 UNTS 171, arts 19, 21 (entered into force 23 March 1976). Available online: https://treaties.un.org/doc/Publication/UNTS/Volume%20999/volume-999-I-14668-English.pdf.
[47] Office of the United Nations High Commissioner for Human Rights, "Business and Human Rights" (2024), online: OHCHR https://www.ohchr.org/en/business-and-human-rights.
[48] Office of the High Commissioner for Human Rights, *Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework* (Geneva: United Nations, 2011), online: OHCHR<https://www.ohchr.org/en/publications/reference-publications/guiding-principles-business-and-human-rights>.

a. Increased Transparency in Algorithmic Moderation

Social media platforms must prioritize transparency in their moderation practices. The UNGPs advocate for mechanisms like independent audits and public disclosure of moderation policies to ensure fairness and equity.[49] While tools like Facebook's Ad Library and Twitter's transparency reports are steps in the right direction, they often lack meaningful stakeholder input and fail to address cultural and geopolitical biases. Collaborating with affected Sikh individuals can help platforms develop culturally informed policies and prevent misrepresentation.[50]

b. Independent Oversight Bodies and Accountability

Governments should establish independent oversight bodies to monitor content moderation and address algorithmic bias. These bodies require diverse representatives from civil society, academia, and marginalized communities to ensure equitable governance. The EU's *Digital Services Act* provides templates for structuring such oversight.[51] Additionally, public databases of moderation precedents, as suggested by Common[52], could improve transparency and procedural fairness while fostering user trust.

Oversight mechanisms must operate within international frameworks like the *International Covenant on Civil and Political Rights* ("ICCPR") to align moderation practices with global human rights standards. International treaties must hold state actors accountable for exploiting algorithms to suppress dissent. Laws like India's Unlawful Activities (Prevention) Act, which broadens the definition of terrorism to suppress dissent, highlight the urgency of enforcing international accountability using international human rights frameworks like the ICCPR.

Binding international agreements are essential to ensuring freedom of expression in digital spaces, protecting democratic discourse, and addressing systemic inequities. To address these challenges, such binding international treaties should enforce accountability for states and corporations. These treaties must prioritize equity and protect marginalized communities, particularly in the Global South, from digital repression. The Sikh American Legal Defense and Education Fund highlight the global impact of India's censorship policies, which influence moderation practices even on U.S.-based platforms.[53] Such spillover effects underscore the need for global frameworks to safeguard digital rights and prevent cross-border suppression.

Call for International Action

To uphold international legal standards and safeguard freedom of expression in digital spaces, the following actions are crucial:

1. Binding Digital Rights Treaties: Nations should develop treaties that extend the protections of the ICCPR to the digital realm, addressing algorithmic moderation and foreign interference. These treaties would ensure that both states and platforms uphold democratic principles and human rights in digital governance.

2. Platform Accountability Frameworks: Social media companies must adhere to international frameworks, such as the UNGPs, ensuring they implement due diligence

[49] Supra note 8
[50] Supra note 10
[51] European Commission, *The Digital Services Act (DSA): Cooperation Mechanism*, online: Digital Strategy https://digital-strategy.ec.europa.eu/en/policies/dsa-cooperation.
[52] Supra note 33
[53] Supra note 12.

processes and independent audits. These measures would promote equitable and transparent practices while mitigating human rights violations.

3. Global Oversight Mechanisms: An independent international oversight body is essential to monitor compliance with digital rights standards, mediate cross-border abuses, and ensure transparency in algorithmic governance.

Implementing these actions could address systemic challenges, protect marginalized voices like pro-Sikh activists, and promote equity in digital spaces. Grounding these measures in established international frameworks reinforces the need for a unified global approach to algorithmic governance and human rights protection.

Conclusion

Social media platforms must transition from tools of digital repression to facilitators of democratic discourse. The suppression of pro-Sikh activism highlights the broader challenges of algorithmic governance, emphasizing the need for transparency, accountability, and equitable solutions. Addressing these challenges is essential to ensuring technology fosters inclusion and equity rather than exacerbating systemic inequities.

As AI technologies evolve, immediate action is critical to prevent these systems from silencing dissent and undermining democratic principles. This paper proposes algorithmic audits, independent oversight bodies, and binding international treaties to offer a practical and scalable framework for creating fairer digital ecosystems.

This is not just a policy concern but a pressing human rights issue. Protecting freedom of expression, particularly for marginalized communities, is vital to upholding democracy in a digital age. Achieving this requires global collaboration, ethical governance, and steadfast accountability from both states and platforms.

The future of digital spaces depends on balancing technological innovation with justice and equity. By taking into considerations the measures proposed in this paper, the international community may ensure that technology amplifies the voices of the marginalized and becomes a tool for progress rather than repression.

References

Books

John Stuart Mill, *On Liberty* (London: John W Parker & Son, 1859).

Noam Chomsky, The Responsibility of Intellectuals (New York: New York Review of Books, 1967).

Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (New York: PublicAffairs, 2019).

Articles

Jasjit Singh, "Narratives in Action: Modelling the Types and Drivers of Sikh Activism in Diaspora" (2020) 11:10 Religions 539.

Reports

Office of the High Commissioner for Human Rights, "Business and Human Rights" (2024), online: OHCHR https://www.ohchr.org/en/business-and-human-rights.

Sikh American Legal Defense and Education Fund, "Written Testimony on Disinformation Nation: Social Media's Role in Promoting Extremist and Misinformation," submitted to the United States House of Representatives, Committee on Energy & Commerce, Subcommittee on Communications and Technology and Subcommittee on Consumer Protection and Commerce (25 March 2021).

World Sikh Organization of Canada, *Enforcing Silence: India's War on Sikh Social Media* (July 2020).

Online: News

Al Jazeera, "In India, Government Now Has Power Over Social Media Content Moderation" (28 October 2022), online: https://www.aljazeera.com/economy/2022/10/28/in-india-govt-now-has-power-over-social-media-content-moderation.

Amnesty International, "Myanmar: Facebook's Systems Promoted Violence Against Rohingya; Meta Owes Reparations – New Report" (29 September 2022), online: Amnesty International https://www.amnesty.org/en/latest/news/2022/09/myanmar-facebooks-systems-promoted-violence-against-rohingya-meta-owes-reparations/.

Amnesty International, "Meta's Failure Contributed to Abuses Against Tigray in Ethiopia" (31 October 2023), online: Amnesty International https://www.amnesty.org/en/latest/news/2023/10/meta-failure-contributed-to-abuses-against-tigray-ethiopia/.

AP News, "Canada-India Diplomatic Tensions: Allegations of Violence and State-Sponsored Actions" (September 2023), online: https://apnews.com/article/canada-india-diplomats-violence-80e9a0d43faa7db99781563434d0f1e0.

BBC News, "Nodeep Kaur: Indian Activist Allegedly Beaten and Sexually Assaulted in Custody" (15 February 2021), online: BBC https://www.bbc.com/news/world-asia-india-56071706.

Business & Human Rights Resource Centre, "Meta responds to allegations of content moderation failures facilitating misinformation, harmful content ahead of Indian elections" (2023), online: https://www.business-humanrights.org/en/latest-news/meta-responds-to-allegations-of-content-moderation-failures-facilitating-misinformation-harmful-content-ahead-of-indian-elections/.

CBC News, "Twitter Blocked Canadian Tweets at India's Request, Including Those by Rupi Kaur and Jagmeet Singh" (2023), online: CBC News https://www.cbc.ca/news/canada/british-columbia/twitter-canada-india-rupi-kaur-jagmeet-singh-1.6787760.

Centre for Information Resilience, "Return of the RealSikhs: The Fake Network Targeting Sikhs Across the World Despite Platform Takedowns" (2023), online: CIR https://www.info-res.org/cir/articles/return-of-the-realsikhs-the-fake-network-targeting-sikhs-across-the-world-despite-platform-takedowns/.

Emily Bembeneck, Rebecca Nissan & Ziad Obermeyer, "To Stop Algorithmic Bias, We First Have to Define It" (21 October 2021), online: Brookings Institution https://www.brookings.edu/articles/to-stop-algorithmic-bias-we-first-have-to-define-it/.

Geo News, "Indian Fake Media Network Targeting Sikhs, Farmers, Separatists Exposed in UK Research" (2021), online: Geo News https://www.geo.tv/latest/384101-indian-fake-media-network-targeting-sikhs-farmers-separatists-exposed-in-uk-research.

Independent, "Jack Dorsey Says Twitter Was Pressured by India During Farmers' Protests" (2023), online: Independent https://www.independent.co.uk/tech/jack-dorsey-twitter-india-protests-b2356403.html.

International Covenant on Civil and Political Rights, 16 December 1966, 999 UNTS 171, arts 19, 21 (entered into force 23 March 1976). Available online: https://treaties.un.org/doc/Publication/UNTS/Volume%20999/volume-999-I-14668-English.pdf.

Jas UK Singh, "This is the reward for speaking the truth, raising human rights and challenging oppression & inequality... #Censorship Well only in India! @jack @Twitter @TwitterIndia #FreeJaggiNow #NeverForget84 #SikhGenocide84 #Sikhs #Truth #Democracy #FreedomOfSpeech #UndueInfluence #Silencing" (18 February 2022, 5:08 PM), online: Twitter https://x.com/UK51NGH/status/1494795952011554819

Jasjit Singh, "Racialisation, 'Religious Violence' and Radicalisation: The Persistence of Narratives of 'Sikh Extremism'" (2020) 46:15 Journal of Ethnic and Migration Studies 3136 at 3147, online: https://doi.org/10.1080/1369183X.2019.1623018.

Jaskaran Sandhu, "India Is Now The World's Largest Electoral Autocracy," Baaz News (22 August 2023), online: https://www.baaznews.org/p/jaskaran-sandhu-india-is-now-the?utm_source=publication-search.

Jasmeen Bassi, "How India Uses Communications to Suppress Farmers and Sikhs" (2021), Baaz News online: https://www.baaznews.org/p/jasmeen-bassi-how-india-uses-communications?utm_source=publication-search.

MacKenzie F. Common, "Fear the Reaper: How Content Moderation Rules Are Enforced on Social Media" (2020) 34:2 Int'l Rev of L, Computers & Tech 126, online: Taylor & Francis https://doi.org/10.1080/13600869.2020.1733762.

National Post, "Canadian Sikh Group Alleges Censorship After Indian Government Asks Twitter to Delete Its Post" (2023), online: National Post https://nationalpost.com/news/canada/canadian-sikh-group-alleges-censorship-after-indian-government-asks-twitter-to-delete-its-post.

Office of the High Commissioner for Human Rights, *Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework* (Geneva: United Nations, 2011), online: OHCHR

https://www.ohchr.org/en/publications/reference-publications/guiding-principles-business-and-human-rights.

Public Inquiry Into Foreign Interference in Federal Electoral Processes and Democratic Institutions. Final Report. Vol 1: Report Summary. Toronto: Privy Council Office, 2025.

Prem Thakker, "Twitter Admits It's Forcing Elon Musk's Tweets Onto Everyone's Timeline" (31 March 2023), online: The New Republic https://newrepublic.com/post/171543/twitter-admits-its-forcing-elon-musk-timeline.

Reporters Without Borders, "Modi Ramps Up Online Censorship in India" (4 September 2023), online: https://rsf.org/en/modi-ramps-online-censorship-india.

Rupi Kaur, "@Twitter has censored Sikh media accounts which have been showcasing the reality of protestors on the ground... this decision—amid an internet ban currently enforced by India—will cost lives. #FarmersProtest #SikhCensorship" (26 January 2021, 3:28 PM), online: Twitter https://x.com/rupikaur_/status/1354164407329153024.

Sikh Coalition, "Holding Social Media Platforms Accountable for Censorship" (2024), online: https://www.sikhcoalition.org/our-work/legal-and-policy/holding-social-media-platforms-accountable-for-censorship/.

Sikh Coalition, *So Many Targets: Contextualizing Modern Indian Transnational Repression Against the Sikh Community* (2024), online: Sikh Coalition https://www.sikhcoalition.org/wp-content/uploads/2024/07/So-Many-Targets-Sikh-Coalition-TNR-Report.pdf

SpringerLink, "Virtually (Non)existent? The Role of Digital Media in Russian LGBTQ+ Activism" (July 2023), online: SpringerLink https://link.springer.com/article/10.1057/s41311-024-00592-7.

Stuart A Thompson, "Elon Musk's X Posts: A Look at Misinformation and Influence" (27 September 2024), online: The New York Times https://www.nytimes.com/2024/09/27/technology/elon-musk-x-posts.html.

The Washington Post, "India's Facebook Problem: Propaganda and Hate Speech Flourish Ahead of Elections" (26 September 2023), online: The Washington Post https://www.washingtonpost.com/world/2023/09/26/india-facebook-propaganda-hate-speech/.

United States, Department of State, 2023 Country Reports on Human Rights Practices: India (2023), online: Department of State https://www.state.gov/reports/2023-country-reports-on-human-rights-practices/india/.

World Sikh Organization, "Criminalizing Dissent in India: A Conversation with Rana Ayyub," YouTube (28 October 2021), online: https://www.youtube.com/watch?v=FUQ8ls_Flb4.

Chapter 12

When Machines Hire: Why Human Oversight and HITL Mechanisms Cannot Solve Bias Embedded in Machine-Learning AI Recruitment Systems

Angeli Patel (she/her)

JD Candidate, Lincoln Alexander School of Law

Abstract

Artificial Intelligence (AI) is rapidly being implemented in hiring processes nationwide, particularly through the use of algorithms to screen resumes, rank applicants, and predict work performance. In response to growing apprehension about AI bias, the use of "human-in-the-loop" (HITL) mechanisms have increasingly gained traction. These approaches aim to combine machine-learning AI models with human oversight, taking the position that human judgment can correct or mitigate bias produced by AI decisions. However, this short commentary argues that, despite the growing reliance on HITL in recruitment practices within Canada's labor market, these systems have not effectively achieved their intended goals of reducing bias. Using a feminist and intersectional theoretical lens, HITL mechanisms are critiqued for perpetuating systemic inequality in hiring practices through the use of biased unrepresentative training data. Systematic and structural amendments, such as improving transparency, conducting regular bias audits, and designing more equitable AI systems, are encouraged. The significance of this analysis lies in its contribution to the broader discourse on the implications of AI in the context of labour and employment. By examining the intersection of AI, bias, and HITL systems, this paper highlights practical recommendations for mitigating systemic inequalities in AI-driven recruitment processes.

Keywords: machine learning, AI, human-in-the-loop, hiring, recruitment

Introduction

Based on 2018 LinkedIn data, approximately 64% of employers use AI tools and data analytics in their recruitment strategies to automate candidate screening and streamline the hiring process.[1] With the explosion in GenAI, a more recent finding showed that six in ten recruiters are optimistic about AI in recruitment.[2] Popular go to models for applicant screening and interviewing are HireVue, SeekOut, interviewAI. Recruiters in Canada are no exemption as a 2023 survey by Indeed revealed that AI in talent acquisition is gaining traction. According to the survey, only 8% of Canadian HR and talent acquisition leaders reported that their teams are not currently utilizing AI tools.[3]

Certainly, most talent recruiters recognize that AI in recruitment brings many challenges, including algorithmic bias. However, they also note that its use in the hiring process will continue to rise notwithstanding. This is due in most part to reciprocal use of AI for job applications by applicants which in turn has spike application volume. One strategy to manage the drawbacks of AI use in hiring is the "human-in-the-loop" (HITL) mechanisms whereby recruiters can combine the efficiency of AI systems with imperative human oversight.

HITL can be broadly defined as the involvement and oversight by human examiners throughout the lifespan of AI-based systems.[4] Despite the regulatory objectives behind this approach, it is clear that HITL does not sufficiently address - and in some cases even exacerbates - the biases inherent in AI recruitment tools, which often favour certain types of candidates over others.[5]

While HITL structures can be undeniably flawed in their execution due to the personal biases of human evaluators, I contend that the root of the problem lies in the biased unrepresentative training data that is used to develop these AI algorithms. This data unconsciously creates and reinforces discriminatory frameworks, making HITL systems ineffective at reducing bias in AI-driven recruitment decisions.

HITL in the Recruitment Process

The recruitment process in Canada is irrefutably tedious and time-consuming, which leads recruiters to spend very little time on the average resume. In fact, a 2012 study found that hiring committees and recruiters spend four to five minutes, on average, assessing a single resume.[6] More than a decade later, with a highly competitive job market and a large volume of applications for each job posting, it is understandable why many companies, hiring committees, and recruiters have turned to AI-driven employment systems to streamline and restructure this practice.

To preserve the efficiency of these AI systems which filter resumes, predict work performance, and rank candidates based on predefined measures such as qualifications and skills, HITL mechanisms have been introduced to address growing concerns about the biased conclusions drawn by these

---

[1] Josephine Yam & Joshua A. Skorburg "From Human Resources to Human Rights: Impact Assessments for Hiring Algorithms" (2021) 23 Ethics and Information Technology 613 [Yam & Skorburg].

[2] LinkedIn Talent Solutions, The Future of Recruiting 2024: AI will supercharge recruiting (21 March 2024) online: LinkedIn <https://business.linkedin.com/content/dam/me/business/en-us/talent-solutions/resources/pdfs/future-of-recruiting-2024.pdf> 4.

[3] Indeed, Canadian HR leaders reveal how AI helps in recruitment and hiring (21 March 2024) online: Indeed <https://ca.indeed.com/leadershiphub/canadian-hr-ai-recruitment-hiring>.

[4] Markus Langer et al. "Effective Human Oversight of AI-Based Systems: A Signal Detection Perspective on the Detection of Inaccurate and Unfair Outputs" (2024) 35:1 Minds and Machines 1 [Langer et al.].

[5] Christopher G. Harris, "Combining Human-in-the-Loop Systems and AI Fairness Toolkits to Reduce Age Bias in AI Job Hiring Algorithms" (2024) University of North Colorado 60 [Harris].

[6] Will Evans, *Keeping an Eye on Recruiter Behavior* (Boston University, 2012), online: https://www.bu.edu/com/files/2018/10/TheLadders-EyeTracking-StudyC2.pdf. [Evans].

techniques.[7] Many researchers have expressed their praise for HITL systems, including scholar Chris Harris who states that "HITL effectively addresses biases […], enhancing fairness and accuracy by combining human and AI capabilities".[8] He argues that HITL provides important oversight and expertise to AI decision making, serving as an essential check and balance to ensure accuracy, fairness, and equity in AI-generated recruitment judgments.[9] This can be supported by a study by Dastin et al., in which revealed that HITL measures significantly decreased bias in the hiring process and, therefore, led to a more equitable selection of candidates.[10]

While I acknowledge that HITL is a progressive instrument in an overall effort to address AI bias in hiring decisions, I argue that HITL is not as effective as it is commonly perceived. Systemic biases continue to persist in the use of AI-driven recruitment systems even with HITL intervention. This is predominantly due to the use of biased training datasets that shape how these systems make evaluations.

Biased Training Data

The limitations of recruitment-based HITL efforts are particularly evident when examined in relation to AI algorithmic design, with one of the most significant issues being biased and unrepresentative training data.[11] This occurs when social biases are unintentionally encoded within the training data of an AI system, resulting in flawed projections about the suitability of a candidate for a particular role.[12] Historical hiring practices, such as those based on gender or ethnicity, are often clear examples of this issue.[13]

If a training dataset contains a disproportionate number of positive markers for men applicants, for example, the AI system may struggle to predict positive markers accurately for women candidates.[14] This issue is further compounded when the data is unrepresentative or imbalanced, failing to appreciate the diversity of the candidate pool.[15] Particular demographic groups, especially women and racial minorities, are often underrepresented in training data, which results in the AI system being more likely to make errors in predicting, rating, or screening for these groups.[16] This is particularly pertinent when an algorithm is trained to prioritize candidates from prestigious universities.[17] This results in the system overlooking qualified candidates who possess equivalent experiences but do not fit into these niche and narrow markers of success.[18] This disproportionately disadvantages low-income women and racial minorities, who often face systemic barriers that limit their access to such institutions, further deepening the inequalities perpetuated by AI-driven hiring systems.[19]

Not only does this issue shed light to feminist and intersectional theoretical lenses but also reinforces the issue of AI failing to fairly represent non-binary and non-white individuals in its

[7] Abdulrahman Wael "The Power of Artificial Intelligence in Recruitment: An Analytical Review of Current AI-Based Recruitment Strategies" (2023) 8:6 International Journal of Professional Business Review 10 [Wael].
[8] Harris, *supra* note 5 at 61.
[9] *Ibid*.
[10] *Ibid* at 62.
[11] Kevin Bauer et al. "Feedback Loops in Machine Learning: A Study on the Interplay of Continuous Updating and Human Discrimination" (2024) 25:4 Journal of the Association for Information Systems 807 [Bauer et. al.]
[12] Wael, *supra* note 7 at 7.
[13] *Ibid* at 11.
[14] Bauer et al., *supra* note 11.
[15] *Ibid*.
[16] *Ibid*.
[17] Wael, *supra* note 7 at 12.
[18] *Ibid*.
[19] Bauer et al., *supra* note 11.

training data.[20] This can be supported by the work of Eleanor Drage and Federica Frabetti who emphasize that AI systems are inherently biased against racialized women, creating harmful exclusionary effects on these individuals and communities.[21]

A practical demonstration of this can be seen in Amazon's failed attempt to implement an AI recruitment tool.[22] The system, which was trained on resumes submitted to the company over a ten-year period, inadvertently learned from the company's gender imbalance.[23] Given the historically low number of women employed at Amazon, as in most technology companies, the algorithm began to identify male-dominated patterns in the data and categorized male candidates as more favorable.[24] By using the results of its own predictions to influence future hiring decisions, the algorithm became trapped in a self-reinforcing cycle of bias against female candidates.[25] This created an unreflective feedback loop that perpetuated gender inequality.[26] Despite efforts to address the issue, Amazon ultimately scrapped the tool, acknowledging that the algorithm's biases could not be eliminated as the underlying predispositions were so deeply enrooted in the system's training data.[27]

This illustration underscores the fundamental problem with AI recruitment tools: if the experiences, skills, and qualifications of women are not adequately integrated in the training data of these AI recruiting systems, how can we expect these AI-driven systems to treat women fairly in the decisions that it produces? While HITL mechanisms are designed to introduce human oversight to correct these issues, they are fundamentally ineffective at addressing the deep-seated biases, lack of diversity, and gaps in representation embedded in the training data. HITL cannot resolve this problem if the very criteria used to assess candidates are flawed from the outset.

HITL as an Illusion of Fairness

I assert that the issue with the effectiveness of HITL mechanisms is that human oversight cannot correct biases that have already been so deeply encoded into the AI hiring algorithm. Although scholars argue that human intervention can potentially adjust certain decisions or outcomes, it is evident that the systemic bias inherent in the training data of these hiring algorithms remains largely intact.[28] AI systems trained on biased data tend to continue to perpetuate these biases, and human intervention after-the-fact is often not enough to overcome the depth of these issues.[29] In fact, scholar Reuben Binns argues that human oversight in hiring decisions can sometimes amplify bias rather than mitigate it, especially when human decision-makers are not adequately trained to recognize and address the biases present in the system.[30] This is reiterated in the work of other scholars, who argue that HITL can become a "feedback loop of data-driven echo chambers" or another opportunity for personal biases to influence the hiring decisions, especially if humans are unaware of how their own biases may support or bolster those encoded in the AI system.[31] As a

[20] Eleanor Drage and Federica Frabetti, "AI that Matters: A Feminist Approach to the Study of Intelligent Machine" in Jude Brown et al eds *Feminist AI: Critical Perspectives on Algorithms, Data, and Intelligent Machines* (Oxford University Press, 2023) 274. [Drage & Frabetti].
[21] *Ibid*.
[22] Maude Lavanchy "Amazon's Sexist Hiring Algorithm Could Still Be Better Than A Human" 2018 International Institute for Management Development 1 [Lavanchy].
[23] *Ibid*
[24] *Ibid*.
[25] *Ibid*.
[26] *Ibid*.
[27] *Ibid*.
[28] Harris, *supra* note 5 at 60.
[29] *Ibid*.
[30] Reuben Binns "Human Judgment in Algorithmic Loops: Individual Justice and Automated Decision-Making" (2020) 16:1 Regulation & Governance 6 [Binns].
[31] Bauer et al., *supra* note 11 at 807.

result, HITL systems may appear to reduce partiality on the surface, but in practice, they reinforce the very patterns of exclusion and bias they were generated to address.[32]

If the underlying instructional data is biased and discriminatory at its core, I contend that the final hiring decision is also innately unfair, regardless of the HITL measures that have applied to reduce it. AI systems are only as unbiased as the data they are trained on - biased training data will always lead to biased results.[33] The quality of the data used to train the algorithms is critical to the effectiveness of its predictive outputs.[34] Without addressing the root cause - the unrepresentative training data itself - HITL remains an insufficient solution to addressing bias, acting as a superficial remedy to a much deeper problem.

Conclusion

HITL mechanisms are built upon training datasets that are often embedded with discriminatory and biased hiring qualities that reproduce systemic biases and prejudices. Without addressing the initial prejudices and limitations in the training data itself, any HITL mechanisms that proceed cannot effectively reduce bias and will continue to act as a false sense of oversight for companies deploying AI-driven algorithms in their recruitment efforts.

For HITL systems to be effective, the training data in which it is constructed upon must be transparent, diverse, and represent a broad range of genders, backgrounds, and skills. Hiring committees and recruiters must conduct audits regularly and stay vigilant and critical of the results produced by these AI systems. Further research on who exactly should be tasked with acting as human overseers in AI-driven hiring tools is warranted.

[32] Binns, *supra* note 30.
[33] Wael, *supra* note 7 at 10.
[34] *Ibid* at 9.

References

Books

Drage, E. and Frabetti, F. "AI that Matters: A Feminist Approach to the Study of Intelligent Machine" in Jude Brown et al eds *Feminist AI: Critical Perspectives on Algorithms, Data, and Intelligent Machines* (Oxford University Press, 2023) 274.

Articles

Bauer K. et al. "Feedback Loops in Machine Learning: A Study on the Interplay of Continuous Updating and Human Discrimination" (2024) 25:4 J*ournal of the Association for Information Systems* 807.

Binns, R. "Human Judgment in Algorithmic Loops: Individual Justice and Automated Decision Making" (2020) 16:1 *Regulation & Governance* 6.

Harris, C. "Combining Human-in-the-Loop Systems and AI Fairness Toolkits to Reduce Age Bias in AI Job Hiring Algorithms" (2024) *University of North Colorado* 60.

Langer, M. et al. "Effective Human Oversight of AI-Based Systems: A Signal Detection Perspective on the Detection of Inaccurate and Unfair Outputs" (2024) 35:1 *Minds and Machines* 1.

Lavanchy, M. "Amazon's Sexist Hiring Algorithm Could Still Be Better Than A Human" 2018 *International Institute for Management Development* 1.

Wael, A. "The Power of Artificial Intelligence in Recruitment: An Analytical Review of Current AI-Based Recruitment Strategies" (2023) 8:6 *International Journal of Professional Business Review* 10.

Yam, J. and Skorburg J. "From Human Resources to Human Rights: Impact Assessments for Hiring Algorithms" (2021) 23 *Ethics and Information Technology* 613.

Online Materials: Websites

Evans, W. "Keeping an Eye on Recruiter Behavior" (2012) *Boston* University online: https://www.bu.edu/com/files/2018/10/TheLadders-EyeTracking-StudyC2.pdf.

Indeed, Canadian HR leaders reveal how AI helps in recruitment and hiring (21 March 2024) online: Indeed <https://ca.indeed.com/leadershiphub/canadian-hr-ai-recruitment-hiring>.

LinkedIn Talent Solutions, The Future of Recruiting 2024: AI will supercharge recruiting (21 March 2024) online: LinkedIn <https://business.linkedin.com/content/dam/me/business/en-us/talent-solutions/resources/pdfs/future-of-recruiting-2024.pdf> 4

Chapter 13

Preventing Racism in Law: The Implementation of AI in Canadian Law using Critical Race Theory

Erin L. Peterson

JD Candidate, Lincoln Alexander School of Law

Abstract

The integration of artificial intelligence (AI) into the Canadian legal system presents significant risks of reinforcing racial bias and systemic discrimination. This paper argues that Critical Race Theory (CRT) must be employed to guide AI regulation in Canada, ensuring its implementation promotes equity rather than exacerbates existing disparities. AI models used in legal settings—such as facial recognition, predictive policing, and recidivism assessment—have demonstrated racial biases that disproportionately harm marginalized communities. Despite ongoing regulatory efforts, Canada lacks a framework that explicitly addresses these risks through a Critical Race lens. By analysing the intersection of CRT, AI, and the Canadian legal system, this paper highlights the necessity of proactive measures to prevent discriminatory outcomes. It calls for a temporary halt on AI deployment in legal settings until CRT-informed regulations are established and enforced. Additionally, it advocates for diverse regulatory bodies, stronger accountability mechanisms, and measures that align the interests of policymakers and marginalized communities. Without deliberate intervention, AI will replicate racial inequalities present in the legal system, highlighting the urgency of adopting a CRT framework to ensure that AI enhances, rather than undermines, substantive equality in Canadian law.

Keywords: Critical race theory, Canadian legal system, artificial intelligence, legal theory

Introduction

The artificial intelligence (AI) industry in Canada has grown significantly, and its use has become widespread across the country. In 2023, a survey across fourteen countries found that 28% of 14,000 users use generative AI at work, with an additional 32% expecting to use it in the near future.[1] Over half of are using the tools without formal employer approval.[2] The prevalence of legal AI systems in Canada has been growing exponentially, with tools such as LexisNexis Legal AI Tools,[3] and with OpenAI's 2022 generative AI model performing better on the U.S. bar exam than about 10% of the human test-takers, improving to 90% the following year.[4] Unregulated generative AI use has been documented in Canadian courts.[5] Concurrently, it is being demonstrated that AI systems are prone to outcomes of racial discrimination,[6] highlighting the risk of implementing AI in the Canadian legal system.

In this short paper, I reflect on the AI from a Critical Race Theory (CRT) vantage and demonstrate that it is imperative the Canadian Government employ CRT when constructing the regulations governing the use of AI in the Canadian legal system. I propose a pause of the use of AI in the Canadian legal system until regulations incorporate a Critical Race lens.

To understand the importance of employing this framework, we must examine the intersection of CRT, AI, and the Canadian legal system. The paper begins in Part II with a brief background on CRT, and its historical intersections with AI and the law. I explain how the employment of this lens in some respects has yielded positive results, and the lack of implementation in others has led to ethical failures. In Part III, I discuss the unregulated misuse of AI in legal systems around the world, pointing to the ethical issues stemming from the use of this technology, and ending with specific Canadian cases. Part IV outlines the steps that Canada is taking to regulate AI, noting aspects of regulation which directly impact the legal system. In Part V, I argue that the historical interactions of CRT with the Canadian legal system and AI, as well as the demonstrated negative implications of the use of AI in law, prove that a lens of CRT must be employed when adopting AI in our legal system to achieve substantive equality. I specifically note how the lens of CRT will impact legislation, painting a picture of how it will look to adopt AI in the Canadian legal system using a Critical Race lens.

Background

Critical Race Theory

Delgado et al. describe the CRT movement as "a collection of activists and scholars engaged in studying and transforming the relationship among race, racism, and power," through a "broader perspective that includes economics, history, setting, group and self-interest, and emotions and the unconscious."[7] A CRT framework can be applied in any field to work towards racial equity and justice, such as public administration, medicine, technology, and law. The basic tenets of CRT

[1] Salesforce, "More than Half of Generative AI Adopters Use Unapproved Tools at Work | Salesforce" (15 November 2023), online (blog): <salesforce.com/news/stories/ai-at-work-research/>.
[2] *Ibid.*
[3] LexisNexis, "LexisNexis Legal AI Tools" (2024), online: <lexisnexis.ca/en-ca/products/legal-ai-tools.page>.
[4] OpenAI, *GPT-4 Technical Report* (OpenAI, 2024).
[5] *Zhang v Chen,* 2024 BCSC 285 [*Zhang*].
[6] Donnesh Dustin Hosseini, "Generative AI: a problematic illustration of the intersections of racialized gender, race, ethnicity | OSF Preprints" (4 November 2024), online (pdf): doi.org/10.31219/osf.io/987ra.
[7] Richard Delgado, *Critical Race Theory: An Introduction*, 4th ed, Critical America (New York: University Press, 2023) at 3.

outline a set of beliefs developed by Critical Race theorists that provide insight on how one would view the world through the lens of CRT.

The tenets that inform how AI should be adopted into the Canadian legal system include:[8]

Racism should be viewed as normal: CRT understands that "racism is ordinary, normal, and embedded in society," and that most people of colour are continually discriminated against in public and private spheres.[9] This is to say, it is the ordinary experience of most people of colour and should be considered a risk in all situations.

Interest-convergence: as demonstrated by Derrick Bell in 'Brown v. Board of Education and the Interest-Convergence Dilemma,' arguing that CRT asserts that the majority will only secure civil rights for minorities when it serves the majority's self-interest.[10]

Intersectionality: the concept coined by Kimberlé Crenshaw, by which everyone has overlapping identities, loyalties, and allegiances which potentially compound to cause greater discrimination.[11] Crenshaw argues that "any analysis that does not take intersectionality into account cannot sufficiently address the particular manner in which Black women are subordinated."[12]

A unique voice of colour: members of racial minorities have a unique perspective and "a presumed competence to speak about race and racism" in ways that the majority group does not.[13] This leads to the conclusion that the self-expressed views of victims of oppression provide essential insight into the nature of the legal system.[14]

Colour-blindness: a racial ideology that proposes the best way to end discrimination is by treating individuals as equal as possible without regard to the racial, cultural, or ethnic background.[15] This ideology works to serve the majority group by allowing them to ignore the inequities, racial disparities, and the history of violence that exist within society.[16]

These tenets shape the framework that I propose be used when implementing AI in the Canadian legal system.

Critical Race Theory in Canadian Law

Scholars have studied the application of CRT in the Canadian legal system, but the failure to apply these findings have led to discriminatory outcomes. In *Critical Race Theory: Racism and the Law*, published in 1999, Carol Aylward points to many significant forms of oppression for Black people and other people of colour in the Canadian legal system.[17] Specifically, Aylward highlights cases of police brutality, racial discrimination, and jury selection. Aylward posits that two steps be

[8] I have chosen five tenets of CRT to focus on in this paper as they are more informative to this issue, but further analysis could be done using Delgado's remaining tenets of "race as a social construct" and "differential racialization."
[9] Delgado, *supra* note 7 at 8.
[10] Derrick A Bell, Jr, "Brown v. Board of Education and the Interest-Convergence Dilemma" (1980) 93:3 Harv L Rev 518—33.
[11] Delgado, *supra* note 7 at 16.
[12]Kimberlé Crenshaw, "Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics" (1989) 1989:1 U Chicago Legal F 139—67.
[13] Delgado, *supra* note 7 at 11.
[14] The Editors of Encyclopaedia Britannica, "Critical race theory - Racism, Oppression, Inequality | Britannica" (15 November 2024), online: <britannica.com/topic/critical-race-theory>.
[15]Monnica T Williams, "Colorblind Ideology Is a Form of Racism | Psychology Today" (7 December 2011), online (blog): <psychologytoday.com/us/blog/culturally-speaking/201112/colorblind-ideology-is-form-racism>.
[16] Kevin Gordon, "Multiculturalism is Better than Colorblindness" (21 November 2013), online (blog): <campkupugani.com/multiculturalism-better-colorblindness>.
[17] Carol A Aylward, Canadian Critical Race Theory: Racism and the Law (Nova Scotia: Fernwood Publishing, 1998) at 134.

followed in a Canadian Critical Race Litigation strategy; the first is to be aware of the social reality of racism in Canadian society and look at the law from the perspective of people of colour and allow that perspective to form our analysis.[18] The second step in a Critical Race Litigation strategy is to "identify when 'race' is an issue."[19] The advocate must locate the problem within the social reality of racism.[20] Decades later, Aylward's recommendations become relevant in the application of AI in the Canadian legal system, which I will explore further in Part V.

There are many reports of racial discrimination in the Canadian judicial system, including findings from the Office of the Correctional Investigator which report:

That the same systemic concerns and barriers identified nearly a decade ago, including discrimination, stereotyping, racial bias and labeling of Black prisoners, remain as pervasive and persistent as before. In fact, the situation for Black people behind bars in Canada today is as bad, and, in some respects, worse than it was in 2013.[21]

This report also states that despite lower rates of reoffending and returning to custody, Black persons are more likely to be assessed as high risk, low motivation, and low reintegration potential.[22] These findings demonstrate the CRT tenet of *interest-convergence*; since working to decrease discrimination is not in the majority's self-interest, the civil rights of minorities have not been prioritized or secured. The demonstrated lack of a CRT lens in the Canadian legal system, along with the negative outcomes, highlight the importance of employing this lens in the future.

Critical Race Theory in AI

AI development has demonstrated a lack of regard for CRT, leading to unethical outcomes. Racism and sexism are embedded into facial detection models, generative AI, and AI in highly sensitive fields such as banking or healthcare.[23] Gebru describes the bias perpetuated by AI predictions as a "runaway feedback loop of increasing the existing marginalisation" caused by "using past data to determine future outcomes."[24] This concern has been recognized globally, with the United Nations Special Rapporteur KP Ashwini, urging "states to regulate these technologies within an approach that recognizes structural racism and is based on key human rights standards."[25] I will highlight key failures of AI models in relation to racial discrimination.

a. Facial Detection Models

Discriminatory outcomes verified in facial detection and recognition models demonstrate a disregard of CRT. For instance, a 2012 study by Brendan Klare et al reviewed three state of the art commercial facial recognition algorithms (Cognitec's FaceVACS v8.2, PittPatt v5.2.2, and Neurotechnology's MegaMatcher v3.1) and found that they all exhibited lower recognition

[18] *Ibid* at 136.
[19] *Ibid*.
[20] *Ibid*.
[21] Office of the Correctional Investigator, Press Release, "Correctional Investigator says Situation for Black People in Canadian Federal Penitentiaries has not Improved Ten Years After Landmark Investigation" (1 November 2022), online: <oci-bec.gc.ca/en/content/correctional-investigator-says-situation-black-people-canadian-federal-penitentiaries-has>.
[22] *Ibid*.
[23] KP Ashwini, Contemporary forms of racism, racial discrimination, xenophobia and related intolerance, GA/HRC/RES/52/36, UNGAOR, Fifty-sixth session, Item 9, A/HRC/56/68 (2024) at para 40; Donald E Bowen III et al, "Measuring and Mitigating Racial Disparities in Large Language Model Mortgage Underwriting" (2024) 4812158 Social Science Research Network at 16.
[24] Timnit Gebru, *Oxford Handbook on AI Ethics Book*, Chapter on Race and Gender (England: Oxford University Press, 2019) at 7.
[25] Ashwini, *supra* note 23 at para 66.

accuracies on females, Black people, and younger subjects aged eighteen to thirty years old.[26] Buolamwini and Gebru later introduced a novel method of intersectional demographic evaluation of face-based gender classification accuracy, allowing for greater insight into discrimination demonstrated by the models. They then used this method to evaluate three more recent facial recognition models, and it was found that all models performed best for lighter individuals and males overall and performed worst for darker females.[27] It has been proven that the demographics of the team building facial recognition algorithms and the demographics of the datasets that they are trained on corelate to the accuracy of the algorithms on individuals of different races.[28] If the development of these models was approached through the lens of CRT, the tenets of a *unique voice of colour* and *racism being viewed as normal* would have identified the likelihood of discriminatory outcomes, and specific interventions could be proactively implemented to combat the discrimination.

b. Generative AI

Generative AI is a term to describe algorithms which generate new content based on input from a user, compared to traditional AI, which classifies inputs or provides recommendations. Generative AI can now be used to create original images, videos, musical competitions, or even to engage in conversation with, after providing a text prompt, or an image, audio, or video file. Like all AI outputs, the generations are based on patterns found in previous data, and most training data for the publicly available generative AI models is from the internet, media, and literature, which are heavily biased.[29]

Luccicio et al introduced a method for exploring the social biases in text to image systems and used this method to evaluate three publicly available systems. Each system was found to consistently under-represent marginalized identities to different extents.[30] There has been separate evaluation of the popular image generation tool DALL-E, which demonstrated racist and gendered representations generated of Black American women.[31] The lack of understanding that *racism should be viewed as normal,* and that the potential harm is *intersectional*, has led to an ignorance of the bias propagated and the discrimination inflicted when a model is trained on uncontrolled datasets.

c. Sensitive (High Risk) Fields

The 2024 report of the UN Special Rapporteur KP Ashwini outlines three areas of healthcare in which AI has demonstrated discriminatory outcomes; health risk scores, disease detection, and AI-enabled medical devices.[32] Racial disparities were worsened in maternal health outcomes in the U.S. by a calculator which predicts the likelihood of successful vaginal birth; it had two race-based correction factors resulting in lower predicted vaginal birth success rates for women of African descent and Hispanic women.[33] Regarding disease detection, AI algorithms for skin cancer

---

[26] Brendan F Klare et al, "Face Recognition Performance: Role of Demographic Information" (2012) 7:6 IEEE TransInformForensic Secur 1789–1801 at 12.
[27] Joy Buolamwini & Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification" (2018) 81:1–15 Proceedings of Machine Learning Research at 12.
[28] Jonathon P Phillips et al, "An Other-Race Effect for Face Recognition Algorithms" (2011) 8:2 ACM Transactions on Applied Perception 1–11 at 8.
[29] Vytenis Kaubrė, "LLM Training Data: The 8 Main Public Data Sources" (27 September 2024), online (blog): <oxylabs.io/blog/llm-training-data>.
[30]Alexandra Sasha Luccioni et al, "Stable Bias: Analyzing Societal Representations in Diffusion Models" (2023) 37th NeurIPS.
[31] Hosseini, *supra* note 6 at 16.
[32] Ashwini, *supra* note 23 at para 40.
[33] Darshali A Vyas et al, "Challenging the Use of Race in the Vaginal Birth after Cesarean Section Calculator" (2019) 29:3 Womens Health Issues 201–204 at 203.

detection have shown poorer performance for those with a darker skin tone.[34] This is caused by the lack of diversity in many data sets used to train these and other AI models.[35] Discriminatory outcomes were demonstrated within AI-enabled medical devices when the use of pulse oximetry devices provided overestimations of the blood oxygen levels of people with darker skin tones.[36]

Moreover, a study which looked at racial disparities in large language models applied to loan underwriting found that applications by a Black borrower are less likely to receive an approval recommendation than otherwise-identical White applicants' applications.[37] These discriminatory outcomes in sensitive fields represent significant repercussions of failing to employ the lens of CRT in the model development, particularly the tenet of *colour-blindness*.

Discriminatory and Unregulated AI in Legal Systems

There have been demonstrated discriminatory outcomes stemming from the use of AI in legal systems around the world. Two ways in which AI is used in law enforcement, security, and the criminal justice system are predictive policing algorithms and recidivism assessment algorithms.[38] There has also been demonstrated misuse of AI within Canadian courts.

a. Predictive Policing

Predictive policing tools are trained on public and personal data and used to make predictions on where, and by whom, future crimes will be committed. This is known to exacerbate over policing of racialized communities, demonstrating the feedback loop that Gebru identified.[39] Racially biased predictions are present within actively used predictive policing tools, caused by biased training data as well as its mathematical oversimplification.[40] These biased predictions lead to increased policing in racialized communities, which leads to an increase in the bias of the data that the model will be retrained on – and the feedback loop continues.

Racial bias within predictive policing algorithms was also proven in a 2022 study highlighting the role of racial bias in the output of predictive policing applications in the U.S., which produce a significantly greater number of predictions towards Black Americans than White Americans.[41]

Another example of predictive policing is live video monitoring, where companies are producing AI systems that use machine vision to perform shoplifting detection on customers in a store.[42] Other facial recognition tools have resulted in six known instances of wrongful arrest in the U.S.,

---

[34] Likhitha Kolla & Ravi B Parikh, "Uses and limitations of artificial intelligence for oncology" (2024) Cancer 130(12), 2101—07.
[35] David Wen et al, "Characteristics of publicly available skin cancer image datasets: a systematic review" (2022) 4:1 The Lancet Digital Health e64–e74 at 71.
[36] Ashwini, *supra* note 23 at para 43.
[37] Bowen, *supra* note 23 at 16.
[38] Tendayi Achiume, Racial Discrimination UN Human Rights Council Special Rapporteur on Contemporary Forms of Racism & Un Secretary-General, "Contemporary forms of racism, racial discrimination, xenophobia and related intolerance: note /: by the Secretary-General" (2019), online: <https://digitallibrary.un.org/record/3827500> at 7.
[39] Alexia Gallon & Emily Liu, "Racism Repeats Itself: AI Racial Bias in Predictive Policing Algorithms" (1 November 2023) at 7 online (pdf):< thinkyou.bayhonors.org/wp-content/uploads/BHS_Papers/2023_Symposium/AlexiaGallon_RacismRepeatsItself.pdf>.
[40] Gallon, *supra* note 39 at 17.
[41] *Ibid.*
[42] Inkryptis AI, "AI Theft Detection | Inkryptis AI", online: <inkryptis.com/ai-theft-detection>.; Veesion, "Our Solution | Veesion" (2024), online: <veesion.io/en/our-solution/>.

with all six victims being Black people.[43] The propagation of discrimination stems from the pre-existing inequality I have described both in AI models and in the current legal system.

b. Recidivism Assessment Algorithms

Recidivism assessment algorithms act by predicting the likelihood of an individual reoffending a crime after being released on bail, sentence, or parole by providing a 'recidivism score.' Courts in the U.S. have used the 'recidivism risk scale' component of the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) to perform recidivism assessments since 2000, influencing pretrial decision-making regarding detention or release and bail setting, determining probation or parole conditions, and the proper placement of offenders in state and federal prisons with appropriate levels of security for over one million offenders.[44]

In *Ewert v Canada*, the Canadian Supreme Court found that some of the recidivism assessments employed by the Correctional Services of Canada may discriminate against Indigenous prisoners, breaching s.24(1) of the *Corrections and Conditional Release Act*.[45] To understand why these instances of algorithmic unfairness occur, we can look to Berk et al, who drew on the existing literature in criminology, computer science, and statistics to examine the fairness and accuracy of criminal justice risk assessments. It was concluded that "there are at least six kinds of fairness, some of which are incompatible with one another and with accuracy," and that, "except in trivial cases, it is impossible to maximize accuracy and fairness at the same time, and impossible simultaneously to satisfy all kinds of fairness."[46] This indicates that if the most valuable goal of these tools is accuracy, fairness will not be maximized.

Canadian Courts

The Canadian Judicial Council's "Guidelines for the Use of Artificial Intelligence" ("GUAI") states that:

> The adoption of AI cannot be a passive or reactive process. Some forms of AI are already embedded in everyday judicial applications for tasks such as translation, grammar checking, speech recognition and legal research. As generative AI becomes more prevalent, it becomes imperative that judges appreciate the implications, limitations, evolving risks, and mitigation strategies associated with its use.[47]

This set of guidelines came after the case of *Zhang v Chen,* which demonstrates a highly publicised misuse of AI in the courts of British Columbia where Ms. Zhang seeks costs against Mr. Chen's counsel, "on the basis that [counsel] inserted into the notice of application two non-existent cases, which were discovered subsequently to have been invented by ChatGPT."[48] Mr. Chen's counsel

---

[43] Alyxaundria Sanford, "Artificial Intelligence is Putting Innocent People at Risk of Being Incarcerated | Innocence Project" (14 February 2024), online: <innocenceproject.org/artificial-intelligence-is-putting-innocent-people-at-risk-of-being-incarcerated/>.
[44] Thomas Blomberg et al, Validation of The COMPAS Risk Assessment Classification Instrument (Florida State University: Center for Criminology and Public Policy Research, 2010) at 43.
[45] *Ewert v Canada,* 2018 2 SCR 165 at para 63.
[46] Richard Berk et al, "Fairness in Criminal Justice Risk Assessments: The State of the Art" (2018) 50:1 Sociological Methods & Research 3.
[47] Martin Felsky & Karen Eltis for the Canadian Judicial Council, Guidelines for the Use of Artificial Intelligence in Canadian Courts (Canadian Judicial Council, 2024) at s 1 [*GUAI*].
[48] *Zhang*, *supra* note 5 at para 2.

was ordered to review all previous filings with the court, and to advice of any use of ChatGPT in the future.

AI is also being integrated into the Canadian immigration system, often covertly, positioning Canada at the forefront of AI-driven immigration and border control technologies.[49] Advanced machine learning algorithms analyse and triage large volumes of applications, detecting potential fraud, while facial recognition kiosks verify identities at entry points.[50] However, these technologies have demonstrated racial bias, as highlighted in *Barre v Canada*, the first Canadian case challenging the use of facial recognition in immigration. The Refugee Protection Division refused to disclose details about the technology used in photo comparisons, raising concerns over transparency.[51] Despite the court confirming the use of facial recognition, these concerns remain unresolved, with cases of AI-driven refugee revocation disproportionately affecting Black individuals and people of colour continuing to emerge.[52] The presence of AI in Canadian courts is already contributing to discriminatory outcomes and is expected to persist.

What Canada is Doing

While our focus has been on the potential issues that arise from adopting AI in the Canadian legal system, Canada has been making progress in the regulation of widespread AI. The Global AI Law and Policy Tracker, released by the International Association of Privacy Professionals, highlights Canada as one of twenty-six jurisdictions of focus making steps in AI governance.[53] Three pieces of regulation that have aspects relating to the adoption of AI in the Canadian legal system are the *Artificial Intelligence and Data Act* (*AIDA)*, the "GUAI", and the "Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems" ("VCC").

The AIDA[54], which has not yet been adopted, notes the potential biased output of AI models based on any prohibited grounds for discrimination, including race and gender, in the *Canadian Human Rights Act* (*CHRA*) as a potential collective harm to Canadians.[55] *AIDA* states that it will protect Canadians from collective harms by requiring businesses conducting regulated activities to proactively assess and mitigate the risk of bias on grounds prohibited in the *CHRA*, as well as the existing option for recourse for discrimination under the *CHRA* or provincial human rights legislation.[56] This would require extensive monitoring to ensure unknown racial discrimination is not occurring, as it is unlikely a victim of AI discrimination would be aware of the occurrence.

The Canadian Judicial Council released the "GUAI" in Canadian courts in September 2024, which states that "any consideration of the use of assistive AI by judges should always be consistent with the core values of the court and judicial ethics." Three of which are equality and impartiality, fairness, and certainty. They propose that this will ensure judges avoid introducing biases, which can inhabit AI outputs, and do not exclude any segments of the population or perpetuate bias against all, including marginalized groups.[57]

---

[49] Gideon Christian, "The New Jim Crow: Unmasking Racial Bias in AI Facial Recognition Technology within the Canadian Immigration System" (2024) 69:3 McGill LJ at 7.
[50] *Ibid* at 11.
[51] *Barre v Canada (Citizenship and Immigration),* 2022 FC 1078.
[52] Gideon, *supra* note 49 at 13.
[53] International Association of Privacy Professionals, "Global AI Law and Policy Tracker" (November 2024), online (pdf): <iapp.org/media/pdf/resource_center/global_ai_law_policy_tracker.pdf> at 2.
[54] Ministry of Innovation, Science and Economic Development Canada, "The Artificial Intelligence and Data Act (AIDA) – Companion document" (13 March 2023), online: <ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document> [*AIDA*].
[55] *Canadian Human Rights Act*, RSC 1985, c H-6.
[56] *Ibid*.
[57] *GUAI*, *supra* note 47 at 7.

The "VCC" was released by the Ministry of Innovation Science and Economic Development and states that all advanced generative AI systems made available for public use "assess and curate datasets used for training to manage data quality and potential biases" and "implement diverse testing methods and measures to assess and mitigate risk of biased output prior to release."[58] This notes some consideration of racial bias in training data for AI models.

The regulatory landscape of AI in Canada has progressed significantly, but has recently slowed due to criticism, controversy, and shifting priorities.[59] The delay in regulation is dangerous in relation to the propagation of bias, as the longer AI is used unmitigated, the larger volume of discriminatory decisions will be made.

AI Implementation through a Critical Race Lens

Examining the intersection of CRT, AI, and the Canadian legal system highlights the potential racially discriminatory outcomes from the implementation of AI in the Canadian legal system. The tenets of CRT can be examined in relation to this implementation to guide the development and deployment of this technology.

The first basic tenet is that *racism should be viewed as normal*, and that continual discrimination of most people of colour is demonstrated by many social indicators. These indicators are then embedded into the data that the AI models are trained on. Without specific intervention these models can produce discriminatory outcomes or predictions, which can be used in high-risk settings to propagate discrimination. This discrimination is amplified by the fact that removing data on the characteristics that one might be discriminated against on the basis of, like race, does not remove the bias from the model's output - it actually amplifies it and makes it impossible to test for.[60] This phenomenon was evident in the U.S. legal system through the COMPAS algorithm's recidivism risk scores, which, despite lacking explicit racial indicators, disproportionately flagged Black defendants as high-risk future offenders—nearly twice as often as White defendants.[61] The second step in Critical Race Litigation proposed by Aylward, that a CRT "analysis requires the advocate to ask the 'race' question and to locate the problem within the social reality of racism,"[62] aligns with this tenet and is applicable in the application of AI in the Canadian legal system. This requires us to understand that the social reality of racism will be engrained in the fabric of the model without specific, intentional steps for prevention.

The second tenet, *interest convergence*, states that the majority will only secure civil rights for minorities when it serves the majority's self-interest.[63] This is a valuable tenet to consider regarding the implementation of AI in law. In this case, the civil rights at risk for minorities include freedom from discriminatory practices under the *CHRA,*[64] which outlines practices that may be the subject

---

[58] Ministry of Innovation, Science and Economic Development Canada, "Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems" (8 November 2024), online: <ised-isde.canada.ca/site/ised/en/voluntary-code-conduct-responsible-development-and-management-advanced-generative-ai-systems>.

[59] Blair Attard-Frost, "The Death of Canada's Artificial Intelligence and Data Act: What Happened, and What's Next for AI Regulation in Canada? | Montreal AI Ethics Institute" (17 January 2025), online: < https://montrealethics.ai/the-death-of-canadas-artificial-intelligence-and-data-act-what-happened-and-whats-next-for-ai-regulation-in-canada/>.

[60] Stephanie Kelley et al, "Antidiscrimination Laws, Artificial Intelligence, and Gender Bias: A Case Study in Nonmortgage Fintech Lending" (2022) 24:6 Manuf Serv Oper Manag 3039–59 at s 8.

[61] Julia Angwin et al, "Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks | ProPublica" (23 May 2016) online: < https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.

[62] Aylward, *supra* note 17 at 84.

[63] Bell, *supra* note 10 at 523.

[64] *Canadian Human Rights Act*, *supra* note 55.

of a complaint. Additional rights which apply to public entities are outlined in section 15(1) of the *Canadian Charter of Rights and Freedoms*, the right to "Equality before and under law and equal protection and benefit of law," which states that "Every individual is equal before and under the law and has the right to the equal protection and equal benefit of the law without discrimination and, in particular, without discrimination based on race..."[65]

The majority's self-interest is the achievement of efficiency and cost savings. The processes required to prevent discrimination in these systems will increase the timeline of implementation and will significantly increase the cost; interests not conducive of producing equitable AI. The fact that the securing of civil rights for minorities when implementing AI in law does not serve the majority's self-interest lends to the belief that it will not be done unless there is careful shaping of regulation which will shape the majority's self-interest to align with the minorities' self-interest.

Individuals interacting with AI in the legal system experience unique discrimination based on their overlapping identities, loyalties, and allegiances. This represents the concept of *intersectionality*, the fourth tenet of CRT, applied to the implementation of AI in the Canadian legal system to "sufficiently address the particular manner in which Black women are subordinated."[66] The notion of *a unique voice of colour* is proposed by the fifth tenant of CRT is that members of racial minorities have a unique perspective and "a presumed competence to speak about race and racism" in ways that the majority group do not. Aylward's finding that CRT "requires that we look at the law from the perspective of people of colour and allow our analysis to be informed by this perspective" stems from this tenet and is applicable to the usage of AI in law.

The concept of *colour-blindness*, in which racial neutrality upholds a system of white supremacy, must be addressed to avoid discrimination in the implementation of AI in the Canadian legal system. Like the second step proposed by Aylward, AI in the framework of Critical Race Litigation must be assessed within the social reality of racism.[67] The tenets of CRT highlight important considerations when implementing AI in the Canadian legal system. Without regard to these considerations, it is likely that this implementation results in discriminatory outcomes.

How this will Look in Canada

I propose that there should be a full pause of the use of AI in the Canadian legal system until the "GUAI" is developed, and the *AIDA* is of force and effect. I follow that the development of the guidelines and the implementation of the *AIDA* be done through a lens of CRT.

It is noted in the *AIDA* Companion Document that the Government will allow "ample time for the ecosystem to adjust to the new framework before enforcement actions are undertaken," amplifying the importance of a halt.[68] The regulations developed under the AIDA should enact adequate punishment that securing civil rights by eliminating discrimination furthers the majority's self-interest. The use of a CRT lens would also urge one to suggest that the proposed prohibitions and enforcement be extended to wilful ignorance to prevent potential racial bias in AI models.

Another important aspect of employing CRT when implementing AI in the Canadian legal system is a focus on the diversity of the team creating the regulations. There should be diversity requirements that are monitored and enforced by the Canadian government. This would apply to

[65] *Canadian Charter of Rights and Freedoms*, s 15, Part 1 of the *Constitution Act, 1982*, being Schedule B to the *Canada Act 1982* (UK), 1982, c 11 a.
[66] Crenshaw, *supra* note 12 at 140.
[67] Aylward, *supra* note 17 at 84.
[68] *AIDA*, *supra* note 54.

the "GUAI," the "VCC," and the *AIDA.* Further work would include analysis of all factors in the implementation of these regulations affected when using a CRT lens.

Conclusion

Whether intended or not, racist outcomes will result if AI is adopted into the Canadian legal system without regard to CRT. As demonstrated in this paper, the lack of a CRT lens has led to discrimination both in the law and within AI, therefore without careful consideration these digressions will be amplified by the integration of AI into the Canadian legal system. This paper argues for a full pause of the use of AI in the Canadian legal system until regulations are in place and enforceable. Additionally, CRT must be employed when developing the "GUAI," the "VCC," and throughout the implementation of the *AIDA* within the Canadian legal system. By focusing on implementing AI with a Critical Race Lens, it is possible that the benefits of AI will outweigh the risks.

References

Primary Sources

Legislation

Canadian Human Rights Act, RSC 1985, c H-6.

Jurisprudence

Barre v Canada (Citizenship and Immigration), 2022 FC 1078.

Ewert v Canada, 2018 2 SCR 165.

Zhang v Chen, *2024 BCSC 285*.

Secondary Sources

Books

Aylward, Carol A, *Canadian Critical Race Theory: Racism and the Law* (Nova Scotia: Fernwood Publishing, 1998).

Delgado, Richard, *Critical race theory: an introduction*, fourth edition, Critical America (New York: University Press, 2023).

Gebru, Timnit, *Oxford Handbook on AI Ethics Book,* Chapter on Race and Gender (England: Oxford University Press, 2019).

Journal Articles

Bell, Derrick A, Jr, "Brown v. Board of Education and the Interest-Convergence Dilemma" (1980) 93:3 Harvard Law Review 518–533.

Berk, Richard et al, "Fairness in Criminal Justice Risk Assessments: The State of the Art" (2018) 50:1 Sociological Methods & Research 3-34.

Bowen, Donald E, III et al, "Measuring and Mitigating Racial Disparities in Large Language Model Mortgage Underwriting" (2024) 4812158 Social Science Research Network.

Buolamwini, Joy & Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification" (2018) 81:1–15 Proceedings of Machine Learning Research.

Christian, Gideon, "The New Jim Crow: Unmasking Racial Bias in AI Facial Recognition Technology within the Canadian Immigration System" (2024) 69:3 McGill LJ at 7.

Crenshaw, Kimberle, "Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics" (1989) 1989:1 The University of Chicago Legal Forum 139-167.

Klare, Brendan F et al, "Face Recognition Performance: Role of Demographic Information" (2012) 7:6 IEEE TransInformForensic Secur 1789–1801.

Kolla, Likhitha & Ravi B Parikh, "Uses and limitations of artificial intelligence for oncology" (2024) Cancer 130(12), 2101–2107.

Luccioni, Alexandra Sasha et al, "Stable Bias: Analyzing Societal Representations in Diffusion Models" (2023) 37th NeurIPS.

Phillips, Jonathon P et al, "An Other-Race Effect for Face Recognition Algorithms" (2011) 8:2 ACM Transactions on Applied Perception 1–11.

Vyas, Darshali A et al, "Challenging the Use of Race in the Vaginal Birth after Cesarean Section Calculator" (2019) 29:3 Womens Health Issues 201–204.

Wen, David et al, "Characteristics of publicly available skin cancer image datasets: a systematic review" (2022) 4:1 The Lancet Digital Health e64–e74.

Reports

Ashwini, KP, Contemporary forms of racism, racial discrimination, xenophobia and related intolerance, GA/HRC/RES/52/36, UNGAOR, Fifty-sixth session, Item 9, A/HRC/56/68 (2024).

Blomberg, Thomas, et al, Validation of The COMPAS Risk Assessment Classification Instrument (Florida State University: Center for Criminology and Public Policy Research, 2010).

Felsky, Martin, PhD, JD & Karen Eltis for the Canadian Judicial Council, Guidelines for the Use of Artificial Intelligence in Canadian Courts (Canadian Judicial Council, 2024).

OpenAI, GPT-4 Technical Report (OpenAI, 2024).

News Releases

Office of the Correctional Investigator, Press Release, "Correctional Investigator says Situation for Black People in Canadian Federal Penitentiaries has not Improved Ten Years After Landmark Investigation" (1 November 2022), online: <oci-bec.gc.ca/en/content/correctional-investigator-says-situation-black-people-canadian-federal-penitentiaries-has>.

Electronic Sources

Angwin, Julia et al, "Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks | ProPublica" (23 May 2016) online: < propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.

Attard-Frost, Blair, "The Death of Canada's Artificial Intelligence and Data Act: What Happened, and What's Next for AI Regulation in Canada? | Montreal AI Ethics Institute" (17 January 2025), online: <montrealethics.ai/the-death-of-canadas-artificial-intelligence-and-data-act-what-happened-and-whats-next-for-ai-regulation-in-canada/>.

Gallon, Alexia & Emily Liu, "Racism Repeats Itself: AI Racial Bias in Predictive Policing Algorithms" (1 November 2023) online (pdf): <thinkyou.bayhonors.org/wp-content/uploads/BHS_Papers/2023_Symposium/AlexiaGallon_RacismRepeatsItself.pdf>

Gordon, Kevin, "Multiculturalism is Better than Colorblindness" (21 November 2013), online (blog):

<campkupugani.com/multiculturalism-better-colorblindness>.

Hosseini, Donnesh Dustin, "Generative AI: a problematic illustration of the intersections of racialized gender, race, ethnicity | OSF Preprints" (4 November 2024), online (pdf): <doi.org/10.31219/osf.io/987ra>.

International Association of Privacy Professionals, "Global AI Law and Policy Tracker" (November 2024), online (pdf): <iapp.org/media/pdf/resource_center/global_ai_law_policy_tracker.pdf>.

Inkryptis AI, "AI Theft Detection | Inkryptis AI", online: <inkryptis.com/ai-theft-detection>.

Kaubrė, Vytenis, "LLM Training Data: The 8 Main Public Data Sources" (27 September 2024), online (blog): <oxylabs.io/blog/llm-training-data>.
LexisNexis, "LexisNexis Legal AI Tools" (2024), online: <lexisnexis.ca/en-ca/products/legal-ai-tools.page>.

Salesforce, "More than Half of Generative AI Adopters Use Unapproved Tools at Work | Salesforce" (15 November 2023), online: <salesforce.com/news/stories/ai-at-work-research/>.

Sanford, Alyxaundria, "Artificial Intelligence is Putting Innocent People at Risk of Being Incarcerated | Innocence Project" (14 February 2024), online: <innocenceproject.org/artificial-intelligence-is-putting-innocent-people-at-risk-of-being-incarcerated/>.

The Editors of Encyclopaedia Britannica, "Critical race theory - Racism, Oppression, Inequality | Britannica" (15 November 2024), online: <britannica.com/topic/critical-race-theory>.

The Ministry of Innovation, Science and Economic Development Canada, "The Artificial Intelligence and Data Act (AIDA) – Companion document" (13 March 2023), online: <ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document>.

The Ministry of Innovation, Science and Economic Development Canada, "Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems" (8 November 2024), online: <ised-isde.canada.ca/site/ised/en/voluntary-code-conduct-responsible-development-and-management-advanced-generative-ai-systems>.

Veesion, "Our Solution | Veesion" (2024), online: <veesion.io/en/our-solution/>.

Williams, Monnica T, PhD, "Colorblind Ideology Is a Form of Racism | Psychology Today" (7 December 2011), online (blog): <psychologytoday.com/us/blog/culturally-speaking/201112/colorblind-ideology-is-form-racism>.

Chapter 14

Critical Approach to AI Development in Sri Lanka: Addressing Tamil Marginalization

Charanija Srirajasingam (She/Her)

JD Candidate, Lincoln Alexander School of Law

Abstract

This paper examines the development of artificial intelligence ("AI") and its governing policies in Sri Lanka through a critical approach lens, focusing on the impact of historical and systemic injustices faced by the Tamil community. Sri Lanka's efforts to build AI technology and data governance frameworks are heavily influenced by a legacy of internal conflict and post-colonial power imbalances. One significant consequence of this legacy is the genocide and hate crimes committed against Sri Lanka's Tamil population, a history that has led to significant distrust towards state-run data collection. Taking a critical approach to this issue reveals how power imbalances and inequalities intersect to shape a nation's data policies in ways that may exclude or marginalize specific groups. This paper argues that, without addressing these systemic issues, Sri Lanka's AI initiatives risk perpetuating exclusionary practices— potentially escalating biased decision-making, privacy violations, and hate crimes against Tamil communities. The resulting AI framework would not only overlook fair representation of the Tamil population but could also enable the surveillance or misuse of data. These challenges prompt debate over whether Sri Lanka's Tamil population would benefit from being in charge of their own data governance, similar to that of Indigenous communities exercising sovereignty through ownership, control, access, and possession ("OCAP") principles. Overall, this paper highlights the importance of inclusive and transparent data governance, setting a precedence for similarly affected communities.

Keywords: Critical Approaches, Artificial Intelligence, Sri Lanka, Tamil Genocide, Data Governance

Introduction[1]

When does the hate begin? Is it when you are a child, walking to school and being raped by state officers?[2] Or is it when you are a baby, ripped from your mother's arms and thrown into a bush as your parents are beaten by Sri Lankan police?[3] For Tamils in Sri Lanka ("SL"), hate is not just a moment—it is a lifetime. It is systemic and sanctioned, fueled by a state that has brutalized my community for generations.[4] As the daughter of Tamil asylum seekers, I see this violence not just as distant history, but as a lived reality. In parallel to these human rights abuses, SL is joining the global trend of integrating Artificial Intelligence ("AI") into governance.

For a community still haunted by genocide, initiatives like AI development do not offer hope—they provoke fear. How can we trust systems built by those who are systematically seeking to erase us? And why should we not expect the same systemic bias to creep into SL's AI framework? These questions form the basis of this paper, which argues that SL's AI development risks reinforcing systemic injustices unless it adopts trauma-informed, inclusive policies that challenge colonial hierarchies and center Tamil data sovereignty.

I will explore how AI can shift from a tool of exclusion and harm to one that advances inclusivity and respects of the Tamils peoples by: (1) critically examining how AI replicates systemic biases rooted in colonial legacies; (2) addressing its misuse as a tool for state surveillance that perpetuates fear and oppression; and (3) incorporating Indigenous data governance frameworks ("IDGF") to challenge systems of oppression that have long defined our reality. While the problems discussed may never be solved, we can imagine a world where technology serves the marginalized—instead of the powerful.

The Illusion of Neutrality

Digital data-driven tools are frequently presented as impartial and objective; however, these assumptions mask the underlying power dynamics present in both their implementation and design.[5] In Margaret Hu's paper, "Critical Data Theory," she reveals that, despite the possibility of amplifying biases, the widespread acceptance of AI governance as superior has allowed for it to be incorporated into decision-making processes with little question.[6] The flaws of AI are visible globally. For example: (1) a Black couple was mistakenly identified as gorillas by Google Photos; (2) racial bias was identified in the COMPAS ("Correctional Offender Management Profiling for Alternative Sanctions") algorithm used in court rulings; and (3) Amazon was found to be using an AI hiring system that disadvantaged female applicants.[7] These are only a few instances of biased results arising from discriminatory training data.[8] Given SL's history of systemic bias and state-led oppression against the Tamil population, these issues are especially concerning. Yet, what is even more alarming is the lack of meaningful discourse surrounding these implications.

---

[1] Note: This paper references news articles and other non-scholarly sources due to the limited availability of academic literature on the topics presented.
[2] See Roy Ratnavel, *#1056: How I Survived War and Found Peace* (Penguin Canada, 2023) which discusses the rape of three Tamil schoolgirls as the reason for the Tamil Tigers ambush on July 23rd, 1983.
[3] Tamil Guardian, "Shock and outrage in Jaffna as Sri Lankan police assault Tamil baby" (10 November 2024), online (news article): <www.tamilguardian.com/content/shock-and-outrage-jaffna-sri-lankan-police-assault-tamil-baby>.
[4] Note: This website provides a detailed overview of the crimes against the Tamil population of Sri Lanka, based on the limited (unbiased) literature available: Pearl Action, "The Tamil Genocide" online: <https://pearlaction.org/tamil-genocide/>.
[5] Margaret Hu, "Critical Data Theory" (2024) 65:4 Wm & Mary L Rev 839 at 850.
[6] *Ibid* at 861.
[7] Ramathi Bandaranayake & Viren Dias, "Towards a Realistic AI Policy for Sri Lanka" (2022) LIRNEasia 1 at 7.
[8] Note: For a brief explanation on biased training data in relation to the Tamil population, see *ibid* at 7-8.

The Committee on Formulating a Strategy for Artificial Intelligence developed an AI strategy titled, "AI Sri Lanka 2028: Sri Lanka's National Strategy on AI", which offers a framework that recognizes bias-related issues and highlights objectives like building public trust through transparency.[9] The Strategy does not, however, address how the historical turmoil between the Tamil and Sinhalese communities will be accounted for in the training data needed to construct AI systems.[10] I have no doubt that many people would likely argue that it is pointless to revisit this history and that we should put the past behind us. However, hate crimes against Tamils are not a thing of the past—they are still happening today.[11] Ignoring the reasons behind SL's current state runs the risk of using AI tools to digitally replicate the same systemic discrimination. To be transparent means to recognize this reality and incorporate it into current discourse, as well as the training data used to build AI systems. As Tamils, are we meant to accept erasure and allow ourselves to be reduced to invisible data points? Or should we keep fighting against systemic injustices that refuse to recognize us as part of humanity?

Post-Coloniality and the Colonial Roots of AI

Post colonial theory offers a critical perspective for evaluating AI systems since this technology is shaped by colonial structures of control and exploitation.[12] SL's colonial history is key to understanding this dynamic. Under British authority, the policy of "divide and rule"[13] amplified tensions between the Sinhalese majority and Tamil minority, with many believing that the latter were favoured over the former. This lie was used as justification for Sinhalese nationalism, which took place following independence.[14] By 1956, political power was unified through policies declaring Sinhalese as the sole official language of SL and prioritizing support for Sinhalese culture and Buddhism.[15] In the present day, AI systems are replicating the dividing techniques that were traditionally used during colonial periods to oppress and control communities by analyzing data to classify, rank, and predict behaviour.[16] If the colonial hierarchies ingrained in AI systems are not critically addressed, this technology risks perpetuating power imbalances that marginalize populations.

Eurocentric behaviour consists of disregarding specific histories and lived experiences to assume a Western universality.[17] To combat this, postcolonial critical sensibilities call for a pluralistic and hybridity of approach to governance, a framework that goes against the "one-size-fits-all" model

---

[9] AI Sri Lanka 2028: Sri Lanka's National Strategy on AI, Draft Strategy for Public Consultation (July 2024) online: <https://mot.gov.lk/assets/files/National_AI_strategy_for_Sri_Lanka 08a88c78d541a0746aeb8c71ed312231.pdf>. p. 26. See also the Committee on Formulating a Strategy for Artificial Intelligence, *Artificial Intelligence in Sri Lanka (White Paper)* (March 2024) https://mot.gov.lk/assets/files/AI%20White%20Paper%20March%202024-c09aa49f7990358ad1442103b804511d.pdf at 3 & 22.

[10] Note: One of the initiatives mentioned in the Sri Lanka AI Strategy" involves launching education campaigns to the Tamil and Sinhalese communities of the state to highlight the potential benefits of AI but scarcely discusses both communities otherwise. See p. 38.

[11] Note: For recent news of hate crimes, see Hannah Ellis-Petersen, Tamils fear prison and torture in Sri Lanka, 13 years after civil war ended (26 March 2022) online: <https://www.theguardian.com/world/2022/mar/26/tamils-fear-prison-and-torture-in-sri-lanka-13-years-after-civil-war-ended> Tamil Guardian *supra* note 3; <https://www.tamilguardian.com/content/tamil-rights-group-highlights-sri-lankas-intimidation-activists-canada-s-foreign>.

[12] See Abeba Birhane, "The Algorithmic Colonization of Africa" (2020) 17:2 SCRIPTed 390. Birhane discussed the concept of "mining" people for data, a colonial practice of treating humans as resources to exploit. See p. 398

[13] Note: See the following source for information on the "divide and rule" strategy in Sri Lanka: Eelapalan, "Did the British Divide & Rule Ceylon?" (23 April 2013), online (blog): <https://sangam.org/british-divide-rule-ceylon/>.

[14] Pearl Action, "History of Sri Lanka" (last visited 29 November 2024) online (blog): <https://pearlaction.org/history-of-sri-lanka/>.

[15] *Ibid*.

[16] Rachel Adams, "Can Artificial Intelligence be Decolonized?" (2021) 46: Interdisciplinary Science Rev 176 at 11.

[17] *Ibid* at 8-9.

used in the Western world.[18] For example, if an AI system was created in SL to identify "high-risk" individuals, it could disproportionately target the Tamil population because of historical and current biases in its training data. A pluralistic approach, however, would require Tamil participation to guide AI development, as such biases may not be obvious to the state—or it may not be a risk they want to consider.

AI as a Tool for State Surveillance

AI has the potential to be used as a weapon by the state, a reality that many do not consider as the technology is being introduced under the pretense of innovation and security.[19] Tamils in SL have experienced their state weaponizing surveillance tactics for years. Between the Sri Lankan armed forces and the Liberation Tigers of Tamils Eelam ("LTTE"),[20] many former LTTE combatants were coerced into becoming informants.[21] Torturing Tamils (those who were part of the LTTE and those who were not) under the powers of the *Prevention of Terrorism Act* became the new norm.[22] As a result, a network of spies was established, consisting of broken-in Tamils who turned on their community as a way to stay alive.[23] Now, with AI unfolding in the National Surveillance State, Tamils have to worry about being transformed into data points used for regulation and control within a system that harnesses an infinite amount of power compared to what was previously experienced.[24]

One example that could be threatened by the misuse of AI is Maveerar Naal, a day close to my heart as it commemorates my uncle's passing in the war and honours hundreds and thousands of those who died in the Tamil liberation struggle.[25] This event, unsurprisingly, is banned by the Sri Lankan state and has been historically (and violently) disrupted by intelligence agencies.[26] State actors photograph participants and later use this as evidence and justification to intimidate and threaten participants—most times going as far as to punish them.[27] What do you think the possibilities could be with the introduction of AI-driven biometric tools that use techniques for advanced data tracking and facial recognition?[28]

These tools could significantly worsen current oppressive practices by providing the state with easier and more efficient methods to monitor and target Tamil individuals, even in spaces or events that are meant to mourn loved ones. Although the war ended in 2009, this technology could still be weaponized to label and surveil Tamils under the vague and unfounded assumption that they

---

[18] *Ibid* at 6 & 8; Birhane, *supra* note 12.
[19] Lakshi Upananda, "Sri Lanka in the Age of Artificial Intelligence: Navigating National Security amid new threats and opportunities" (27 May 2024), online (blog): <https://www.inss.lk/index.php?id=651>.
[20] Note: See the following source for information on the Liberation Tigers of Tamil Eelam: South Asia Terrorism Portion, "Liberation Tigers of Tamil Eelam (LTTE) Sri Lanka" (last visited 8 March 2025), online: <https://www.satp.org/satporgtp/countries/srilanka/terroristoutfits/ltte.htm>.
[21] Immigration and Refugee Board of Canada, "Sri Lanka: Treatment of suspected members or supporters of the Liberation Tigers of Tamil Eelam (LTTE)" (11 February 2015), online: <https://www.irb-cisr.gc.ca/en/country-information/rir/Pages/index.aspx?doc=455729&pls=1>.
[22] *Ibid*.
[23] *Ibid*.
[24] Hu, *supra* note 5 at 872-873. See also, Matthew Tokson, The Authoritarian Risks of AI Surveillance (May 1, 2025) online: Lawfare <https://www.lawfaremedia.org/article/the-authoritarian-risks-of-ai-surveillance>
[25] Tamil Guardian, "Reclaiming the right to mourn" (27 November 2023), online (news article): < https://www.tamilguardian.com/content/reclaiming-right-mourn; Note: See cited source for an explanation on Maveerar Naal.
[26] *Ibid*.
[27] *Ibid*.
[28] Note: See Linnet Taylor, "What is data justice? The case for connecting digital rights and freedoms globally" (2017) 4:2 Big Data & Society 1 for information on India's Aadhaar, the world's largest biometric database which disproportionately disadvantages the poorest (as correlation).

are against the current state.[29] So yes, state surveillance is not a new issue; however, the use of AI risks worsening the effects of this practice under the pretense of national security. If the goal is to make this technology more trustworthy, it needs to be thoroughly examined for its implications in the state that it is being integrated into before it creates a culture of fear.

Incorporating Indigenous Frameworks for Tamil Data Sovereignty[30]

Ownership, control access, and possession ("OCAP") are principles developed by Canada's First Nations to emphasize that Indigenous communities should fully own the data they collect, taking away any unwanted involvement from external state actors.[31] They should choose how their data is collected, analyzed, and processed, which allows them to set rules for how it is governed.[32] Through meaningful engagement and maintaining possession of their data, they can access and interact with it effectively while preventing the state from withholding or altering information.[33] For Tamils in SL, applying these principles would ensure that data of significant importance to their current realities are not manipulated or erased to serve state narratives.

"Tamil-led AI governance frameworks" has a nice ring to it, but what could it actually entail? Let us look at India's Traditional Knowledge Library—an initiative that was created to prevent the appropriation and erasure of Indigenous medicinal knowledge.[34] The implementation of this framework is relevant in SL given its historical attempts to erase Tamil heritage, as seen in the burning of the Jaffna Public Library in 1981, a deliberate act to destroy records of central Tamil history and identity.[35] OCAP principles could prevent similar erasures in the digital age by keeping data intact for future generations. Unfortunately, SL is not off to a great start with its current approach to AI development. The government has appointed a committee under the Presidential Secretariat to lead AI advancements, yet there is no indication that Tamil voices are being included as a key stakeholder.[36] How are Tamils meant to rely on the country's state leaders, given its history of state-led violence? From 2004 to 2015, former president Mahinda Rajapaksa committed atrocious war crimes during the final stages of the civil war.[37] More recently, his brother, Gotabaya Rajapaksa, plunged the country into an economic and political crisis during his reign from 2019 to 2022.[38] The Rajapaksa clan let their hate for Tamils fuel the systemic injustices against the population—how can we believe that Tamil voices will be heard and integrated under the current Presidential Secretariat? The simple answer is that we cannot. Without incorporating OCAP principles into SL's AI models, they risk perpetuating discrimination and making hate crimes more efficient—there is no evidence that shows otherwise.

---

[29] Note: Immigration and Refugee Board of Canada, *supra* note 21 refers to surveillance tactics used during the war.
[30]Chidi Oguamanam "Indigenous Peoples, Data Sovereignty, and Self-Determination: Current Realities and Imperative" (2020) 26 African J Information & Communication 1 at 4.
[31] Brian Schnarch, "Ownership, Control, Access, and Possession (OCAP) or Self-Determination Applied to Research" (2004) 1:1 Intl J Indigenous Health 80 at 81.
[32] *Ibid*; Note: The First Nations Regional Health Survey in Canada is an example of the OCAP principles being applied, as the data collected and analyzed is to be used by the First Nations themselves, allowing them to advocate for better health services. See Oguamanam, *supra* note 30 at 7-8.
[33] Schnarch, *supra* note 31 at 81.
[34] Oguamanam, *supra* note 30 at 7.
[35] Tamil Guardian, "History in flames: remembering the burning of Jaffna Library" (31 May 2021), online (news article): <https://www.tamilguardian.com/content/history-flames-remembering-burning-jaffna-library-0>.
[36] UNDP Sri Lanka, "Stepping into the Age of Artificial Intelligence" (29 March 2023), online: <https://www.undp.org/srilanka/press-releases/stepping-age-artificial-intelligence>.
[37] Tamil Guardian, "War is never the solution, says the war criminal" (17 October 2023), online (news article): https://www.tamilguardian.com/content/war-never-solution-says-war-criminal.
[38] BBC, "Sri Lanka: Why is the country in an economic crisis?" (29 March 2023), online (news article): https://www.bbc.com/news/world-61028138.

The Collective Benefit, Authority to Control, Responsibility, and Ethics ("CARE") is another set of principles that expand on ethical data practices.[39] I like to think of "free, prior, and informed consent" as an umbrella term for CARE, as First Nations in Canada have established governance models that require any use of their data by external consumers to align with Indigenous values and objectives.[40] Tamil communities could adopt similar principles to ensure AI systems in SL recognize Tamils as part of the population without perpetuating biases. A potential application could involve an independent database of evidence documenting war crimes and state-sanctioned violence to provide accountability while ensuring these records are preserved or accessible for the future.

Another application could involve documenting land claims using evidence-based maps and historical records to address the displacement and dispossession of Tamils during and after the war.[41] This would ensure that land restitution claims continue to be made with the confidence of being backed by informed and accurate data instead of state narratives that prioritize their own agendas.[42] Overall, the work-in-progress of IDGF offers a positive perspective on the changes that could be made in SL, where AI could become a tool of empowerment for the Tamil population.

Conclusion

This paper has argued that SL's plan to integrate AI into its governance risks perpetuating systemic injustices unless trauma-informed, inclusive policies are implemented to address Tamil data sovereignty and colonialism. Through a number of critical approaches and indigenous principles it becomes clear that AI in SL could either reinforce historical patterns of oppression or serve as a tool for reconciliation and justice. In a country still dealing with the trauma from its colonial and post-colonial histories, the choice lies in how this technology is designed and governed. The first step towards utopia for SL is to embrace uncertainty; not knowing if hope lies ahead but choosing to be positive as a deliberate act of resistance against what we have come to see as unchangeable.[43] Perhaps the true calling of AI is not merely to advance technology but for its capacity to redefine power.

[39] Oguamanam, *supra* note 30 at 13.
[40] *Ibid* at 10.
[41] Note: See the following links for information on the displacement and dispossession of Tamils: Minority Rights Group, "Tamils in Sri Lanka" (last updated March 2024), online: <https://minorityrights.org/communities/tamils/>; iDMC, "Sri Lanka: A hidden displacement crisis" (31 October 2012), online: <https://www.internal-displacement.org/publications/sri-lanka-a-hidden-displacement-crisis/>.
[42] Elizabeth Fraser, Frederic Mousseau & Anuradha Mittal, *Justice Denied: A Reality Check on Resettlement, Demilitarization, and Reconciliation in Sri Lanka* (Oakland, CA: The Oakland Institute, 2017) at 4; Note: Cited source states that much of the dispossessed land in Sri Lanka is being prioritized for military use and luxury resorts, to align with state objectives.
[43] Guillaume, Chevillon, "The Queer Algorithm" (2024), ESSEC Business School at pg 9.

References

Journal Articles

Adams, Rachel, "Can Artificial Intelligence be Decolonized?" (2021) 46: Interdisciplinary Science Reviews 176.

Bandaranayake, Ramathi & Viren, Dias "Towards a Realistic AI Policy for Sri Lanka" (2022) LIRNEasia.

Birhane, Abeba, "The Algorithmic Colonization of Africa" (2020) 17:2 SCRIPTed 390.

Chevillon, Guillaume, "The Queer Algorithm" (2024), ESSEC Business School.

Hu, Margaret, "Critical Data Theory", Symposium Article (2024) 65:4 Wm & Mary L Rev 839.

Oguamanam, Chidi, "Indigenous Peoples, Data Sovereignty, and Self-Determination: Current Realities and Imperative" (2020) 26 Afr J Inf & Comm 1.

Schnarch, Brian, "Ownership, Control, Access, and Possession (OCAP) or Self-Determination Applied to Research" (2004) 1:1 International Journal of Indigenous Health.

Taylor, Linnet, "What is data justice? The case for connecting digital rights and freedoms globally" (2017) 4:2 Big Data & Society 1.

Online: Websites

BBC, "Sri Lanka: Why is the country in an economic crisis?" (29 March 2023), online (news article): <https://www.bbc.com/> https://www.bbc.com/news/world-61028138.

Hannah Ellis-Petersen, Tamils fear prison and torture in Sri Lanka, 13 years after civil war ended (26 March 2022) online: <https://www.theguardian.com/world/2022/mar/26/tamils-fear-prison-and-torture-in-sri-lanka-13-years-after-civil-war-ended

Immigration and Refugee Board of Canada, "Sri Lanka: Treatment of suspected members or supporters of the Liberation Tigers of Tamil Eelam (LTTE)" (11 February 2015), online: https://www.irb-cisr.gc.ca/en/pages/index.aspx https://www.irb-cisr.gc.ca/en/country-information/rir/Pages/index.aspx?doc=455729&pls=1

Pearl Action, "History of Sri Lanka" (last visited 29 November 2024) online (blog): <pearlaction.org> https://pearlaction.org/history-of-sri-lanka/.

Tamil Guardian, "History in flames: remembering the burning of Jaffna Library" (31 May 2021), online (news article): <tamilguardian.com> https://www.tamilguardian.com/content/history-flames-remembering-burning-jaffna-library-0.

Tamil Guardian, "Reclaiming the right to mourn" (27 November 2023), online (news article): <tamilguardian.com> https://www.tamilguardian.com/content/reclaiming-right-mourn.

Tamil Guardian, "Shock and outrage in Jaffna as Sri Lankan police assault Tamil baby" (10 November 2024), online (news article): <tamilguardian.com> https://www.tamilguardian.com/content/shock-and-outrage-jaffna-sri-lankan-police-assault-tamil baby.

Tamil Guardian, "War is never the solution, says the war criminal" (17 October 2023), online (news article): <tamilguardian.com> https://www.tamilguardian.com/content/war-never-solution-says-war-criminal.

UNDP Sri Lanka, "Stepping into the Age of Artificial Intelligence" (29 March 2023), online https://www.undp.org/srilanka https://www.undp.org/srilanka/press-releases/stepping-age-artificial-intelligence

Upananda, Lakshi, "Sri Lanka in the Age of Artificial Intelligence: Navigating National Security amid new threats and opportunities" (27 May 2024), online (blog): https://www.inss.lk/index.php?id=651.

Policy Work & Reports

AI Sri Lanka 2028: Sri Lanka's National Strategy on AI, Draft Strategy for Public Consultation (July 2024)

Committee on Formulating a Strategy for Artificial Intelligence, "Artificial Intelligence in Sri Lanka" *(White Paper)* (March 2024).

Fraser, Elizabeth, Frederic, Mousseau & Anuradha, Mittal "Justice Denied: A Reality Check on Resettlement, Demilitarization, and Reconciliation in Sri Lanka" (2017) The Oakland Institute.

Chapter 15

Algorithmic Equity in Justice: Reducing Racial and Gender Biases in AI Sentencing Models

Helen Zhang

JD Candidate, Lincoln Alexander School of Law

Abstract

Current Artificial Intelligence (AI) systems often perpetuate historical biases, including racism and sexism, which can influence sentencing practices. This paper will explore methods for training AI to mitigate these inequities in judicial outcomes. Bias has been encoded in AI through their human creators. AI often draws from historical crime data, which reflects longstanding disparities in policing and judicial practices, particularly against Black and Indigenous communities. The biased data can result in a cycle where individuals from these communities are more likely to be considered high-risk offenders, resulting in harsher sentencing outcomes. Developing bias detection and correction algorithms within the AI model can help monitor and flag biased predictions or decision patterns that depart from different demographic groups. These algorithms are designed to recognize and rectify disparate impacts to meet equity standards set by the government. Developers must regularly adjust AI models to consider changing societal patterns and data evolution. Regularly re-evaluating AI outputs for patterns of bias allows for prompt modifications to prevent perpetuating prediction inequities. By prioritizing bias detection and regular monitoring, the courts can create an AI model for judicial settings that recognizes systemic inequality and fosters fairer outcomes.

Keywords: algorithmic sentencing, individualized sentencing, algorithmic bias

Introduction

Artificial Intelligence (AI) generally refers to machines that can replicate certain aspects of human intelligence, such as pattern recognition, decision-making, and prediction. Various sectors use AI for functions such as hiring employees, risk assessment, and criminal sentencing.[1] AI relies on algorithms that process large datasets, adapting and improving with new data. The data that technology, including AI, relies on reflects and reproduces the prevailing social dynamics at the time of collection and encodes bias from its human creators. This reliance makes the machines more prone to perpetuating biases, discrimination, and inequities.[2]

The biased data can significantly shape sentencing practices. AI often draws from historical crime data, which reflects long-standing disparities in policing and judicial practices, particularly against Black and Indigenous communities.[3] The continual use of biased data can result in a cycle where individuals from vulnerable communities are more likely to be considered high-risk offenders.[4] As a result, members from these communities may face harsher sentencing outcomes. This paper will explore methods for training AI to mitigate these inequities in judicial outcomes and discuss the complex results of using historical data for sentencing.

AI as a Sentencing Tool

Governments have been adapting to technological advancements by increasingly introducing AI as a sentencing tool.[5] Risk assessment tools aim to predict the likelihood of an individual re-offending by analyzing various factors such as criminal history, the severity of the current charges before the court, and social and economic circumstances.[6] These systems then generate a risk score, which judges and decision-makers can consider during the sentencing procedure.[7] In Canada, correctional facilities use several algorithmic tools like Custody Rating Scale (CRS), which assesses the security classification of federally sentenced individuals. This system does not use AI, but it relies on information such as age, previous convictions, substance use, risk to public safety and others to make a decision.[8]

Proponents of AI systems argue that introducing them into the risk assessment tools would enhance sentencing fairness and efficiency.[9] The literature suggests that risk assessment tools are generally more accurate in predicting re-offence and recidivism than unregulated human judgment.[10] Furthermore, adequately designed AI algorithms may eliminate unconscious biases that influence human decision-making. AI systems operate solely based on the variables programmed into them, potentially leading to fairer sentencing outcomes. AI can also ensure

---

[1] Carmina Ravanera & Sarah Kaplan, *An Equity Lens on Artificial Intelligence* (Toronto: University of Toronto Institute for Gender and the Economy, 2021) at 2.
[2] Drew Roselli, Jeanna Matthews, and Nisha Talagala, "Managing Bias in AI" in Companion Proceedings of the 2019 World Wide Web Conference (San Francisco, CA USA, May 2019) 539 – 544.
[3] Julia Angwin et al, "Machine Bias" (May 23, 2016) online: Pro Publica <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
[4] Ibid
[5] For a detailed description and analysis of recent AI powered sentencing algorithms in different countries, see Damilola Awotula, Indicium ex Machina: Unstructured Sentencing and Disparate Outcomes in Canada (LLM Thesis, Dalhousie University, Schulich School of Law, 2023) [Unpublished]. Chapter 2, 13 – 36.
[6] Mirko Bagaric et al, "The Solution to the Pervasive Bias and Discrimination in the Criminal Justice System: Transparent and Fair Artificial Intelligence" (2022) 59 Am Crim L Rev 95.
[7] *Ibid*.
[8] Kelly S Montford & Kelly Hannah-Moffat, "The veneers of empiricism: Gender, race and prison classification" (2021) 59: 101475 Aggression and Violent Behavior 2. See also Ewert v Canada, 2018 SCC 30 where the SCC considered the validity of using a risk assessment tool in the context of an Aboriginal offender by the Canadian Correctional Service.
[9] Supra note 5 at 63 – 64.
[10] Bagaric et al, *supra* note 2.

greater consistency in sentencing by applying the same criteria and weightings to all offenders. AI can quickly process vast amounts of data, which enables more efficient sentencing decisions compared to human processing methods, which are time-consuming.[11]

With increasing research in favour of introducing AI into sentencing tools, more risk assessment tools are being rapidly introduced into practice to reduce bias and improve efficiency. However, some systems, such as the CRS, have been criticized for being based on narrow data. It primarily records the experiences of male offenders, which is not an accurate reflection of the risk factors of women.[12] As a result, women - particularly Indigenous women - are often over-classified in the penal system. This highlights some of the most significant flaws of using AI in sentencing tools.

A Critical Race Perspective of Data

AI systems are designed and built by a workforce that is not representative of society, and these AI systems can perpetuate and amplify existing gender biases. An illustration of this is the underrepresentation of women in the AI workforce. It reflects the persistent gender gap of women in Science, Technology, Engineering and Mathematics (STEM) fields. The gap is rooted in stereotypes that associate scientific and technical jobs with masculinity, discouraging women from pursuing careers in these areas.[13] This association is particularly evident in the shift in computer programming from a historically female-dominated field to a male-centric one as it gained prestige and financial rewards.[14]

This lack of diversity in AI development has significant consequences as it leads to the creation of AI systems that are inherently biased against women.[15] The choices behind what and how AI systems are created can impose a narrow set of interpretations and worldviews on the data. If primarily white men are setting AI agendas, then the supposedly "neutral" technology is likely to be embedded with masculine preferences. This creates a "feedback loop," where the under-representation of women in AI development results in biased systems, which further perpetuate gender inequalities.[16] This bias manifest itself in various ways, including the development of algorithms that discriminate against women in hiring, facial recognition software that struggles to identify women accurately, and voice assistants that reinforce gender stereotypes.[17]

The bias encoded in AI by its human creators also negatively affects Black and Indigenous communities as it draws from crime data which reflects historical injustices. AI systems inherently reflect the societal values and inequalities of the environments in which they are created. This means that if AI is developed predominantly by individuals from a specific demographic, it will likely inherit and perpetuate their biases, even if unintentionally.[18]

Law enforcement and the justice system have been introducing AI systems to assist with policing. In theory, this would help improve efficiency and reduce manual labour. However, the caveat is that AI systems analyze historical crime data, which are shaped by societal biases. The biases can

---

[11] *Ibid*
[12] Montford & Hannah-Moffat, *supra* note 4.
[13] Judy Wajcman & Erin Young, "Feminism Confronts AI: The Gender Relations of Digitalisation" in Jude Browne et al eds, Feminist AI: Critical Perspectives on Algorithms, Data, and Intelligent Machines (Oxford, UK: Oxford University Press, 2023) at 47–64.
[14] *Ibid*.
[15] *Ibid*.
[16] *Ibid*.
[17] *Ibid*.
[18] Ravanera & Kaplan, *supra* note 1.

be embedded int AI systems through training data and reflect the disparities in policing and judicial practices.[19] Historical crime data are rooted in systemic racism and the over-policing of these communities. AI systems learn from this data, identifying patterns that reflect racial disparities. This can lead to situations where individuals from Black and Indigenous communities are more likely to be labelled as high-risk offenders by AI tools.[20] It perpetuates the cycle of inequality and can lead to harsher sentencing outcomes, which further marginalize vulnerable communities.

Predictive policing technologies are designed to estimate criminal probabilities. They take existing police data and train algorithms to make predictions based on historical trends.[21] However, they rely on historical crime data that are rooted in racist practices. This often leads to the over-policing of Black and Indigenous neighbourhoods. Predictive policing reinforces the perception of these communities as high-crime areas, leading to increased surveillance and harsher law enforcement practices.

As Ruha Benjamin explains in her book, The New Jim Code, seemingly neutral and objective technologies, can inadvertently perpetuate racial discrimination.[22] Risk assessment tools are intended to assess the likelihood of an individual becoming a victim or perpetrator of violence. However, biased data can also influence them and potentially contribute to discriminatory outcomes. By relying on biased data and operating within existing societal inequalities, these technologies can reinforce and even worsen racial disparities, particularly in areas like law enforcement and criminal justice.

Bias Detection and Correction Algorithms to Monitor and Flag Biased Predictions

Highlighting bias in current data in AI systems helps governments understand AI's limitations—but are there ways to create a more equitable algorithm? AI learns by analyzing large datasets. These datasets come from various sources, such as books, photos, health data, government records, and social media profiles. Algorithms then mine the data, identify patterns, and make predictions, reflecting societal biases from available and missing data. For example, AI systems trained to recognize gender may fail to accurately identify or may misgender transgender and non-binary individuals because they lack sufficient data on these groups. People without online histories, often those coming from low-income or racialized communities, may not be included in AI training data, leading to biased or incomplete outcomes.[23]

One method of rectifying the biased data is to require developers to create a system that detects and corrects bias within AI models.[24] These systems would monitor and flag biased predictions or decision patterns that may impact different demographic groups. Governments should establish standards for designing algorithms that recognize and rectify disparate impacts to meet existing equity standards. Furthermore, developers must regularly adjust AI models to reflect changing societal patterns and data evolution. Regular re-evaluation of AI outputs for patterns of bias allows for prompt modifications to prevent perpetuating prediction inequity. By prioritizing bias detection

[19] Ruha Benjamin, *Race After Technology: Abolitionist Tools for the New Jim Code* (Cambridge, UK: Polity Press, 2019).
[20] *Ibid*.
[21] Kerry McInerney, "Coding 'Carnal Knowledge' into Carceral Systems: A Feminist Abolitionist Approach to Predictive Policing" in Jude Browne et al, eds, *Feminist AI: Critical Perspectives on Algorithms, Data, and Intelligent Machines* (Oxford, UK: Oxford University Press, 2023) 101-118.
[22] Ruha Benjamin, *supra* note 14. See also Gideon Christian, The New Jim Crow - Unmasking Racial Bias in AI Facial Recognition Technology within the Canadian Immigration System, (2023) 69:4 McGill Law Journal 441.
[23] Ravanera & Kaplan, *supra* note 1.
[24] Martin, Julia L. "Unmasking Bias: Investigating Strategies for Minimizing Discrimination in AI Models" (Undergraduate Thesis, Dartmouth College, 2023) at 28.

and regular monitoring, governments can create an AI model for judicial settings that recognizes systemic inequality and fosters fairer outcomes.[25]

Governments must develop a robust legal framework addressing AI's ethical and procedural challenges. This framework should ensure transparency, accountability, and safeguards to protect individual rights and liberties.[26] Transparency and accountability ensure that AI algorithms are accessible and understandable; this allows independent authorities to inspect the algorithms.

Drawbacks of the Bias Recognition System

Fundamentally, we are asking AI to learn how to sentence individuals using historical data that we know is biased. Without intervention, this positive feedback will only perpetuate inequalities. Conversely, consistent intervention may unduly influence data, such that it reflects sentencing not based on precedent but on a political ideal that is unrepresentative of reality.

Addressing bias in AI requires acknowledging the historical and societal contexts that shape data, actively working to dismantle systemic racism, and ensuring diversity and inclusion in the AI workforce. "Fixing" algorithms is insufficient; addressing bias requires a fundamental shift in how we think about and approach technology. We must acknowledge the societal inequalities that shape its development and take concrete steps to ensure diversity and inclusion in the field. This includes challenging racial injustices and gender stereotypes, as well as creating a more welcoming and supportive work environment for diversity in tech.

Governments must carefully craft sentencing tools to prevent the recreation of existing biases. The focus should be on avoiding the inclusion of factors that can act as proxies for discriminatory characteristics like race or socioeconomic status.[27] For instance, the CRS considers factors like substance use and criminal history, which can be influenced by systemic inequalities and over-policing in marginalized communities and lead to biased outcomes.[28] Understanding and controlling these proxy variables can help minimize the potential for biased outcomes.

Addressing Cultural Impacts is Needed Before We Introduce AI into Sentencing

While technological advancements in AI offer potential solutions, biased algorithms are rooted in the more significant systemic issue of flawed data and a lack of objectivity and neutrality within the law. Simply correcting data or improving the predictive capacity of an AI will not solve the bias problem. The issue is not the technology itself but the system in which it operates. This system, including the legal framework, perpetuates bias because it assumes objectivity and neutrality in data and human actors. This flawed assumption codifies existing social prejudices into technical systems, making detecting and dismantling these biases even harder. Focusing solely on technological solutions will not address the underlying societal and systemic issues to lead to genuinely equitable and just AI systems.

Conclusion

[25] Michael Purcell & Mathew Zaia, "Prediction, Prevention and Proof: Artificial Intelligence and Peace Bonds in Canada" (2020) 98:3 Can Bar Rev 515.
[26] *Ibid*.
[27] Mirko Bagaric et al, "The Solution to the Pervasive Bias and Discrimination in the Criminal Justice System: Transparent and Fair Artificial Intelligence" (2022) 59:1 Am Crim L Rev 95.
[28] Kelly Struthers Montford & Kelly Hannah-Moffat, "The veneers of empiricism: Gender, race and prison classification" (2021) 59 Aggression and Violent Behavior (The Classification of Crime: Theory, Research, and Practice) 101475.

Consider this question: Can addressing technology alone resolve the biases it inherits? Achieving equitable outcomes may require not just technological adjustments but also systemic societal change. We must understand how seemingly neutral factors can act as proxies for discriminatory variables like race or gender from data, leading to indirect discrimination. Addressing bias in data requires a comprehensive strategy that considers the social and historical contexts in which data is generated and used, ensuring that AI systems are developed and deployed fairly and equitably, especially in high stake activities like criminal sentencing.

References

Jurisprudence

Ewert v Canada, 2018 SCC 30

Journal Articles

Bagaric, Mirko et al. "The Solution to the Pervasive Bias and Discrimination in the Criminal Justice System: Transparent and Fair Artificial Intelligence" (2022) 59 Am Crim L Rev 95.

Benjamin, Ruha, *Race After Technology: Abolitionist Tools for the New Jim Code* (Cambridge, UK: Polity Press, 2019).

Christian, Gideon, The New Jim Crow - Unmasking Racial Bias in AI Facial Recognition Technology within the Canadian Immigration System, (2023) 69:4 McGill Law Journal 441.

Drew Roselli, Jeanna Matthews, and Nisha Talagala, "Managing Bias in AI" in Companion Proceedings of the 2019 World Wide Web Conference (San Francisco, CA USA, May 2019) 539 – 544.

Montford, Kelly S & Kelly Hannah-Moffat, "The veneers of empiricism: Gender, race and prison classification" (2021) 59 Aggression and Violent Behavior 101475.

Purcell, Michael & Mathew Zaia, "Prediction, Prevention and Proof: Artificial Intelligence and Peace Bonds in Canada" (2020) 98:3 Can Bar Rev 515.

Ravanera, Carmina & Sarah Kaplan, *An Equity Lens on Artificial Intelligence* (Toronto: University of Toronto Institute for Gender and the Economy, 2021).

Book Chapter

McInerney, Kerry, "Coding 'Carnal Knowledge' into Carceral Systems: A Feminist Abolitionist Approach to Predictive Policing" in Jude Browne et al eds, *Feminist AI: Critical Perspectives on Algorithms, Data, and Intelligent Machines* (Oxford, UK: Oxford University Press, 2023) 101.

Wajcman, Judy & Erin Young, "Feminism Confronts AI: The Gender Relations of Digitalisation" in Jude Browne et al eds, *Feminist AI: Critical Perspectives on Algorithms, Data, and Intelligent Machines* (Oxford, UK: Oxford University Press, 2023) 47.

Online: Websites

Angwin, Julia et al, "Machine Bias" (May 23, 2016) online: Pro Publica <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

Theses

Awotula, Damilola, "Indicium ex Machina: Unstructured Sentencing and Disparate Outcomes in Canada" (LLM Thesis, Dalhousie University, Schulich School of Law, 2023) [Unpublished]. See chapter 2, 13 – 36.

Martin, Julia L. "Unmasking Bias: Investigating Strategies for Minimizing Discrimination in AI Models" (Undergraduate Thesis, Dartmouth College, 2023). [Unpublished].

The Transnational Technology Law Review